공학전문석사 학위 연구보고서

비지도 단어 임베딩 기반 토픽 모델링 상세 주제 변화 탐지

– Unsupervised Word Embedding Based Topic Modeling Extracts Latent Biomedical Knowledge from Korean Gov. Research Proposals –

2020년 2월

서울대학교 공학전문대학원

응용공학과 응용공학전공

안 병 은

비지도 단어 임베딩 기반 토픽 모델링 상세 주제 변화 탐지

– Unsupervised Word Embedding Based Topic Modeling Extracts Latent Biomedical Knowledge from Korean Gov. Research Proposals –

지도 교수 김 용 대

이 프로젝트 리포트를 공학전문석사 학위 연구보고서로 제출함

2020년 2월

서울대학교 공학전문대학원

응용공학과 응용공학전공

안 병 은

안병은의 공학전문석사 학위 연구보고서를 인준함

2020년 2월

위 원 장 구 윤 모 (인)

위 원 김 용 대 (인)

위 원 곽 우 영 (인)

국문초록

문자는 정보의 기록과 전달에 가장 효율적인 방법 중 하나이다. 특히 학문적 지식은 문자로 서술되는 경우가 압도적으로 높으며, 최신의 성과는 논문의 형태로 작성 시점을 포함하여 공유된다. 하지만, 정보의 양이 많아 짐에 따라, 효율적인 분석이 필요로 하게 되었으며, 문자의 경우 비구조화된 데이터로 그 의미를 찾아내기 어렵다. 기존에는 표식이 있는 대량의 문서를 기준으로 데이터 분류를 하는 지도학습기반 분석이 되었지만 [1],[2] 이는 변해가는 상황을 반영하여 의미를 도출하기에는 한계가 있다.

본 연구는 비지도 학습 기반의 단어 임베딩을 통해 대규모 문서의 주요 주제와 시간에 따른 상세 주제 변화의 분석을 목적으로 한다. 이를 위해 2006년부터 2017년까지 정부에서 진행된, 바이오 신약개발 연구 과제를 자연어 군집화 방법인 토픽 모델링 방법으로 분석하였다. 전처리 과정 중 전문용어 인식률을 높이기 위해 NPMI를 적용하여, 바이오산업에 특화된 고유명사, 합성명사를 추출하였고, 토픽 모델링은 비지도 기계학습 방법인 LDA(Latent Dirichlet Allocation)으로 문서 내의 거시적인 주제의 변화를 탐색하였다. 이후, 기존 LDA와 단어 임베딩 방법이 결합한 LDA2vec (15만 건의 바이오 기술 관련된 문서로 학습된 단어 임베딩(GloVe) 기반 준 지도 토픽 모델링)을 활용하여, 시간에 따른 신약 연구의 주제뿐만 아니라, 기존의 방법에서 나아간, 주제 내의 단어와 단어의 관계를 탐지하고자 하였다. 이를 통해 특정 단어가 주어졌을 때, 기간별 주제 간의 유사성을 파악하고, 그 유사 주제 안에서 특정 단어 주변의 단어의 움직임을 분석하였다. 이는 주제 내에서의 단어 간의 영향력을 고려할 수 있게 되어, 상세 내용 분석 가능성을 제공한다.

본 연구로 미래의 잠재적 정보의 흐름을 기존 데이터 기반으로 추출할 수 있는 가능성을 보았다. 이는 대량의 학술 문헌 정보에서 학습된 주제 벡터와 단어 벡터를 통해 시간에 따른 상세 주제 변화의 정보를 제공할 수 있다는 점에서, 추후 연구 방향성 설정 및 다양한 자연어 분석이 필요 한 분야에 활용될 것이다.

주요어 : 비지도 토픽 모델링, 단어 임베딩, LDA2Vec, 신약 개발, 자연어 처리

학 번 : 2018-29527

목 차

표 목차

그림 목차

제 1 장 서론

제 1 절 연구 동기

문자의 사용은 인류에게 지식 전달의 효율성을 가져다주었다. 더욱이 3차 산업혁명(디지털 혁명)을 기반으로 야기된, 4차 산업혁명은 2018년 1일 기준 약 2.5 (250경) quintillion bytes[3]의 데이터가 동영상, 사진, 음성 등 다양한 형태로 생산 확산되고 있다. 양적인 데이터 증가 이후 대량의 데이터에서 의미 있는 정보 추출과 같이 질적인 분석의 중요성이 커지고 있는데, 특히 자연어 처리(Natural Language Processing)는 인간이 사용하는 언어를 컴퓨터가 이해하고 분석 처리할 수 있도록 하는 기계 학습의 한 분야로서 문자 데이터 분석에 활용되고 있다. 더욱이 현시대에도 아카데미아에서의 주요 지식은 문자(논문)의 형태로 기록 공유되고 있고, 기업의 전략, 다양한 정보, 및 일상 의사소통 또한, 매체의 변화만 있을 뿐 문자의 형태로 기록, 전달되고 있어, 자연어처리의 중요성은 증가하고 있다. 실례로 기업에서는 데이터 기반 비즈니스 인텔리전트 도출을 위해 자연어 처리 기법을 이용한 소비자의 의견을 분석하거나, 음성 데이터를 문자 데이터로 변형하여[4] 효율적인 의사결정에 활용하고 있다. 또한, 의료 분야에서는 병원이 보유한 EMR(Electronic Medical Record) 환자 정보에 존재하는 정형 데이터 (생체 데이터) 등에 대한 활용도는 높지만, 의사의 진단 의견이 문자로 기록된 비정형 데이터는 분석의 어려움이 있어 자연어처리를 통한 이해를 통해 환자 치료에 도움이 되어야 하는 분야이다. [5] 이에 자연어처리는 지식의 효율적인 활용을 위한 필수의 기술이며, 최근 자연어처리는 2000년 중반 Hinton 교수의 심층신경망으로[6] 시작된 딥러닝의 비약적인 발전과 함께 이미지 분석을 넘어 자연어처리에서도 괄목할 만한 성능을 보인다. 그중 지난 2018년 10월 구글에서 소개한 BERT (Bidirectional Encoder Representations from Transformers)[7]는 11개의 자연어 테스트(SQuAD, 스탠퍼드의 질문 응답 데이터 등)에서 인간의 능력을 뛰어넘는 성능을 보여주었다. 또한, 특정 분야의 문자 데이터를 활용하여, 비지도 학습을 기반 단어 임베딩을 이용 필요 정보를 분석, 예측하는 연구들도 존재한다. 영화 순위 예측[8], 사이언스 문서에서의

특정 물질 조합의 변화 예측[9] 등이 여기에 속한다. 이처럼 자연어 처리 기법은 다양한 분야에서 활발하게 사용되고 있다.

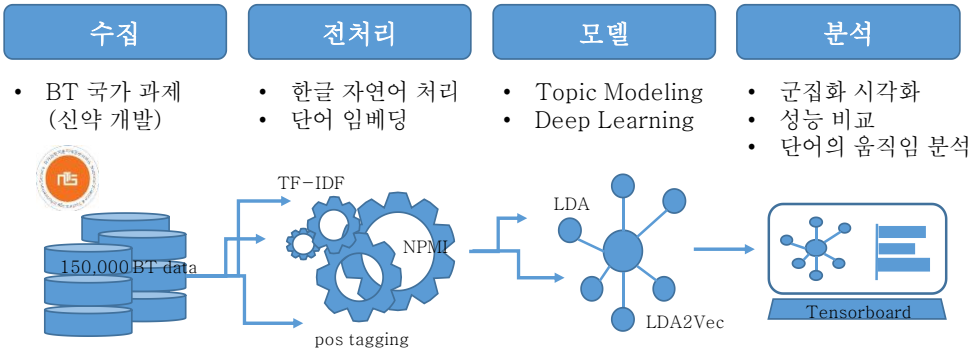
서두에 언급한 것과 같이, 학제별 최신 지식은 논문으로 공유 기록이 되고 있다. 특히 지식 집약적인 산업에서 논문 연구 주제의 추의 분석을 통해 연구의 방향성 설정은 매우 중요하다. 이중 바이오 및 신약개발 분야는 연구 집약적인 산업으로 세계 바이오 헬스 시장 규모는 2024년 2조 6,000억 달러로 예상되며, 이는 반도체, 자동차, 화학제품의 총 시장 규모보다 크다. [10] 바이오 헬스 시장의 신약개발 분야는 그중에서도 대표적인 자본과 시간이 필요한 산업이다. 반면 신약후보 물질 약 7,500종에서 전임상 진행률은 3.3% 정도이며, 임상시험은 0.07%, 최종 승인은 0.01%로 매우 낮다. [11] 따라서, 신약개발의 현 진행 상황 및 방향성에 대한 분석이 중요하다. 이에 본 연구에서는 대한민국 연구 재단에 2006년부터 2017년까지 제안된 BT(Biotechnology) 분야 중 신약 연구 개발 제안서를 전통적 토픽 모델링 기법인 LDA (Latent Dirichlet Allocation)를 통해 1차 분석 진행 후 LDA에 단어 임베딩 기술을 접목한 LDA2vec을 한글에 적용하여 분석 방법의 특징과 신약개발의 주제를 기간별로 탐색하고자 한다. 나아가 특정 약제나 기술에 대한 배경지식이 부족하더라도, 최신 자연어 기법들을 통하여, 신약 연구 분야의 연구 현황과 단기적 개발 방향을 예측하고자 한다.

제 2 절 연구 내용

본 연구에서는 2006년부터 2017년까지 한국 연구 재단에 제안된 이공계 6개 주요 기술 (BT, CT, ET, IT, NT, ST) 연구 제안서 중, BT을 두 가지의 토픽 모델링 기법을 활용하여, 신약개발 분야의 상세 주제를 탐색하고자 한다.

한글 데이터로 구성된 본 제안서 분석을 위해 자연어 처리 모델을 구축하였다. 전문가 도움 없이도, 문서 내의 중요 단어가 분리되고자 하였으며, 제안서의 문서적 특성을 활용하여, 토픽 모델링의 성능을 높이고자 하였다. 이를 위해 3개년을 하나의 기간으로 고려하여 2006년부터 2017년까지를 총 4개의 구간으로 분리하였다. LDA를 통해 문서 내에서의 주요 주제에 대한 생성, 소멸 그리고 비중을 분석하여, 지난 12년간의 신약개발 연구 추이를 분석하고자 하였다. 나아가 문서 벡터로 분석된 토픽 모델링에 단어의 벡터 정보까지 결합하여, 전문지식이 필요한 분야에 더욱더 상세한 분석이 이루어지도록 하였다. 이를 위해 자연어 처리 딥러닝을 활용하여, 단어 임베딩 벡터가 학습되게 하였고, 준지도 학습 구조로 LDA2vec을 실행하였다. 학습된 단어들과 주제의 군집화 성능을 시각화하고, 시간별로 타깃 단어의 변화를 비교하여, 주제의 흐름 분석에 도움이 되는 추가 정보를 제공하고자 하였다.

데이터 획득부터 전처리, 데이터 분석 목적에 맞는 알고리즘 선택과 최신 연구 알고리즘에 관한 연구, 끝으로 결과에 대한 분석으로, 데이터에서 의미 있는 정보를 추출하고자 함에 있다.



[그림 1-1] 토픽 모델링을 통한 자연어 처리 분석 파이프라인

제 3 절 논문 구성

본 연구보고서는 총 5개 장으로 구성되어 있으며, 각 장은 다음과 같은 내용을 포함하고 있다.

1장에서는 자연어 처리와 토픽 모델링을 활용하여 상세 주제 분석을 하는 연구 수행 동기와 연구의 방향성에 관해 설명한다.

2장에서는 본 연구에 필요한 배경 기술을 소개하고, 선행 연구에 대한 조사 결과를 요약하여, 현 학계의 연구 방향을 확인한다.

3장에서는 자연어 전처리 과정과 전처리 성능을 높이기 위한 방법을 실행하고, 토픽 모델링 분석 파이프라인 구현 환경조건을 설명한다.

4장에서는 전처리 성능 향상 확인과, 두 토픽 모델링의 실행 및 수행 결과와 상세 분석을 실행한다.

5장에서는 연구 성과에 대해 요약하고, 추후 후속 연구의 방향과 활용 방안에 대해 논의한다.

제 2 장 배경 지식과 관련 연구

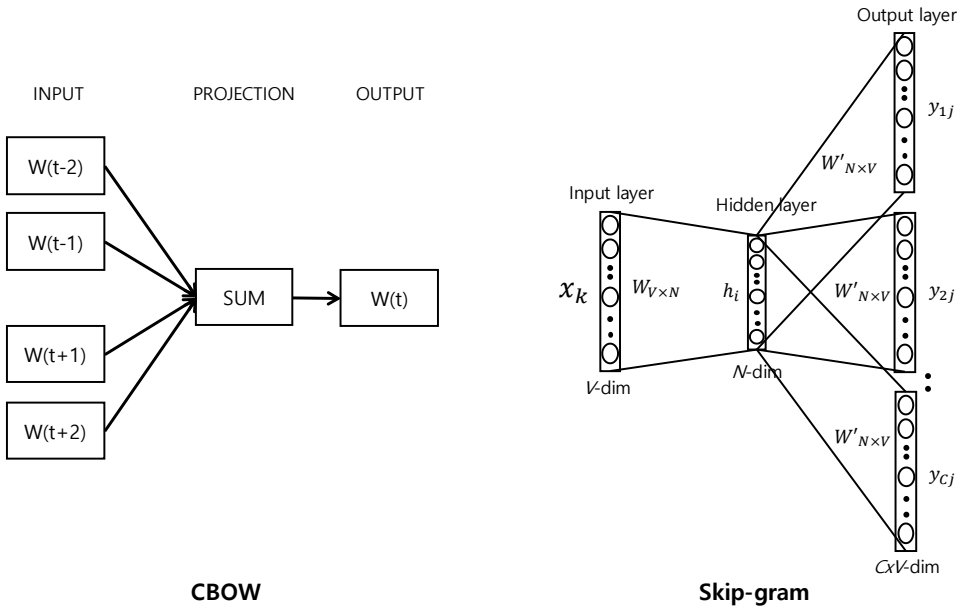
제 1 절 Word Embedding

자연어를 활용하기 위한 기계 학습 방법에서 단어를 지정된 차원의 연속 벡터로 표현하고자 하는 연구는 오래된 역사가 있다. [12, 13] 그중 신경망 통해 학습의 모델을 가시화시켜준 사례는 neural network language model (NNLM)이다. [14]얕은 feedforward deep learning model을 활용하여 단어 벡터를 학습하였다. 초기의 모델로 연산량이 많았고, 학습량이 많아질수록 정보의 손실이라는 단점이 있었다. 이후 NNLM의 일부 한계를 극복한 순환 신경망 Recurrent Neural Net[15]이 순서가 포함된 데이터에서 괄목할 만한 성과를 이루며, 단어 임베딩과, 신경망의 결합은 기계 번역, 문장 생성, 문서 분류 등에 활용되어 자연어이해를 돕고 있다. 분포 가설(Distributional Hypothesis), 같은 문맥에 나타난 단어들은 비슷한 의미를 지니고 있다, [16]에 기반한 단어 임베딩은 크게 1) 데이터 전체의 행렬 인수분해를 통한 차원 축소하는 방법과 2) 문맥창을 통한 데이터의 지역적 정보를 통해 학습하는 방식을 활용한다. [17] 첫 번째 대표적 방식은 LSA(Latent Semantic Analysis)(Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)와 HPCA(Hellinger Principal Component Analysis)(Lebret & Collobert, 2013) 등으로 대표적인 데이터 차원 축소 방식을 따른다. 두 번째는 deep learning의 발전 이후 활발하게 사용되는 방식으로, Word2vec (Google)을 시작으로 GloVe (Standford Univ.), Fasttext (Facebook) 등이 주요 IT 회사에서 개발되었다. 본 연구 또한 문맥 창을 이용한 단어 임베딩 방식을 토픽 모델링과 결합 하여, 신약개발 문서의 분석을 진행하고자 한다.

1. Word2vec

2013년 문맥 창을 기반으로 Google에 의해 제안된 단어 임베딩 기법이다.[18] CBOW와 Skip-gram의 두 가지 방식이 그림 2-1과 같이 소개되었고, 중심 단어로 주변 단어를 예측하는 방식인 Skip-gram이 더 뛰어난 성능을 보이는 것으로 알려져 있다. [19] Skip-gram

의 학습은 아래 식(1)(softmax function)을 최대화하는 방향으로 진행한다.

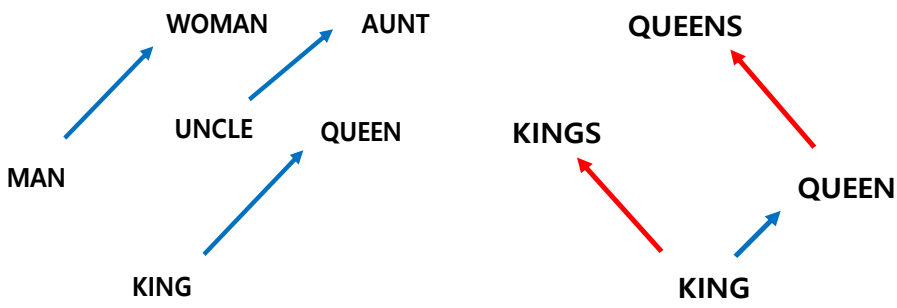


[그림 2-1] Word2Vec의 CBOW와 Skip-gram 학습 방법 [18]

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W} exp(u_w^T v_c)} \quad (1)$$

중심단어(c)가 주어졌을 때, 주변 단어(o)이 나타날 확률을 계산하며,[20] 이를 최대화를 시키기 위해서는 분자의 값이 커야 하는데, 이는 중심어와 주변 단어가 내적 값을 높이는 방향으로 학습을 하며 단어간 내적값이 큰 단어 벡터는 유사도가 높다고 판단할 수 있다. (u는 W'의 열벡터, v는 W의 행벡터) 또한, 분모의 값은 작아야 하는데, 전체 output 벡터와 중심 단어의 내적은 문맥 창에 있는 단어보다 상대적으로 출현 빈도가 낮기에 분모를 감소시키는 효과가 있다. 본 연산 비용은 말뭉치 전체 단어의 수와 연관이 있기에 negative sampling을 사용하여 연산 값을 축소하며 근사 학습을 시킨다. 성공적으로 학습이 진행되었을 경우, 그림 2-2와 같이 남자와 여자, 삼촌과 이모, 왕과 여왕과 같이 유사한 경향성의 단어들이 위치가 같은 방향성을 띠게 되며, 각각 학습된 벡터를 활용해, king에서 man을 빼고, woman을 더하면, queen이란 결과

를 단어 임베딩 정보를 활용하여 얻을 수 있다[21].



[그림 2-2] 남녀 관련 단어와 복수 형태의 명사 단어 임베딩 [21]

2. GloVe

2014년 스탠퍼드 NLP 연구소에서 단어의 벡터 표현을 위해 만든 비지도 학습 알고리즘이다.[17] Word2Vec는 중심어 주변 윈도우 크기 내의 단어들을 고려하여 학습하기에 코퍼스(문자 데이터) 전체의 정보를 반영하지 못하는 단점과 잠재의미분석(LSA)의 경우 단어의 출현 빈도의 통계적인 정보를 차원 축소(SVD)를 통해 분석할 때 단어 간의 관계를 유추하지 못하는 어려움이 있다. GloVe는 Word2Vec과 LSA의 알고리즘을 모두 사용하여, 전체 코퍼스의 정보를 반영한 임베딩 벡터를 만들고자 하였다.

$$J = \sum_{i,j=1}^{V} f(X_{ij})\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - logX_{ij}\right)^2 \qquad (1)$$

Loss Function(J)(1) 을 최소화하는 방향으로 학습을 하게 된다. X_{ij}는 두 단어의 동시 등장 확률(co-occurrence)이다. $w_i, \tilde{w}_j$는 각각의 임베딩된 단어로 $w_i, \tilde{w}_j, b_i, \tilde{b}_j$와 $\log X_{ij}$의 차이가 축소되는 방향으로 학습한다. (V:전체 코퍼스 단어 수, $b_i, \tilde{b}_j$: 조건 성립을 위한 상수항 대체, $f(X_{ij})$: 출현 빈도수 가중치)

3. Fasttext

Facebook에서 2016년 소개한 단어 임베딩 기법으로, 기존 형태학(morphology)를 고려하지 못했던 임베딩 기법들(Word2vec, GloVe)의 한계를 극복한 방법이다. [22] 기본적인 skip-gram의 모델을 따르지만, 기존 띄어쓰기 및 스탬핑의 전처리 기준으로 학습을 하여, 코퍼스가 포함한 단어 이외의 단어에 대해서는 처리를 할 수 없었다. 하지만, Fasttext는 문자를 n-gram로 나누고, 이를 학습한 n-gram의 합으로 표현하여 코퍼스에 포함되지 않은 단어도 유추할 수 있다.

$$s(w,c) = \sum_{g \in \mathcal{G}_w} Z_g^{\intercal} V_c \qquad (1)$$

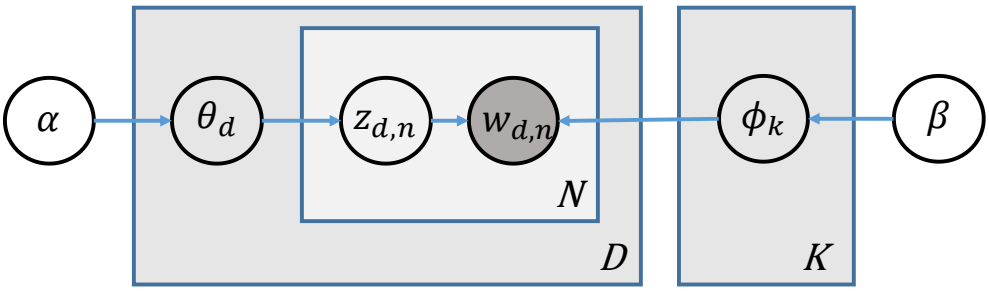
특히 Score Function(1)의 경우 단순 n-gram으로 나뉜 단어뿐만 아니라 추가로 3~6개의 n-gram을(Z) 실시하여 단어 자체도 함께 학습한다. 이는 <Where> 안의 <her> 와 실 단어인 her를 구분하여 학습하는 방법이다.

제 2 절 Topic Modeling

토픽 모델링은 데이터 마이닝 기법의 하나로, 대량의 비정형화된 데이터에서 유의미한 주제(토픽)들을 추출해주는 알고리즘이다. [23] 이는 고차원의 데이터에서 효율적인 차원 축소를 통해 주요 정보는 보전하며, 요약, 추론을 할 수 있게 돕는다. 토픽 모델링 이전의 문서 내 중요 정보를 추출 방법들이 존재하였는데, 출현 시간의 순서대로 보면 대표적으로 TF-IDF(Term Frequency - Inverse Document Frequency) [24]가 있다. TF-IDF는 특정 문서 내에서 어떤 단어가 대표성을 띠는지를 알 수 있는 기법으로, 단어의 빈도와 역 문서 빈도의 곱으로 계산된다. 이후 LSI(latent semantic indexing) [25] 는 단어빈도에서 나아가 문서의 유사도를 계산하고자 하였다. SVD (singular value decomposition)를 통해 단어를 선정하였고, 이는 동의어, 다의어에 대한 특성을 표현할 수 있었다. 하지만 maximum likelihood나 bayesian method 대비 장점에 대해 찾을 수 없었다. 중요한 발전은 Hoffman[26]으로 부터 시작되었는데, 문서 내의 한 단어는 하나의 주제로부터 생성이 되었고, 문서의 다른 단어는 다른 주제로부터 생성이 되었다는 가정하여, 지정된 주제 수의 확률 분포를 줄일 수 있었다. 이를 기반으로 LDA (latent dirichlet allocation)가 2003년 소개되었다.

1. LDA(Latent Dirichlet Allocation)

LDA는 디리클레 분포를 이용한 베이지안 추론으로 주어진 데이터의 주제를 탐색하는 방법이다. LDA는 문서 분석에도 사용되는데, 특히 대량의 문서에서 사람이 해석할 수 있을 만한 단어들의 조합을 제공하여, 주제를 추측해 볼 수 있다. LDA는 다음과 같은 가정을 한다. 하나의 문헌은 여러 개의 주제로 구성되어 있다. 하나의 주제는 여러 개의 단어로 이루어진다. 그리고 각각의 단어는 연구자가 지정한 K개의 주제 중 하나에 포함된다는 3가지의 가정으로 시작된다. LDA의 아키텍처는 다음과 같다. 그림 2-3의 동그라미는 변수, 네모 칸은 반복하여 변수를 업데이트한다. 화살표는 조건과 변수들을 이어주는 역할을 하며, K는 주제의 수, D는 문서의 개수, N은 d번째 문서의 단어의 수를 표현한다. α, β 는 디리클레 분포를 결정하는 하이퍼파라메터다.



[그림 2-3] LDA 모델 아키텍처 1

여기서 관찰 가능한 변수는, $w_{d,n}$, d번째 문서의 n번째 단어가 유일하다. 이후 LDA의 주제별 단어 선택은 하기와 같이 진행된다.

1. 주제 k = 1 to K
   (a) 주제 선정 $\phi_k \sim Dir(\beta)$
2. 총 문서 D에서 각각의 d 문서
   (a) 주제 분포 선택 $\theta_d \sim Dir(\alpha)$
   (b) 각각의 단어의 인덱스 n (from 1 to N_d)
      i. 주제 선택 $z_n \sim \text{Categorical}(\theta_d)$
      i. 단어 선택 $w_n \sim \text{Categorical}(\phi_{k_n})$

변수 추론의 경우 LDA[27] 논문에서는 Variational EM을 사용하여, 추론을 진행하였다. 하지만 논문에서도 언급되었듯이, 라플라스 근사법, 마르코프 연쇄 몬테카를로 방법 등이 사용 가능하다고 하였다.

LDA는 사람이 해석할 수 있는 결과를 제공하지만 반대로 한계도 있다. 특히 문서에서의 주제의 수를 연구자가 임의로 지정해야 하고, 본 알고리즘을 대량의 문서에 적용하였을 때, 그룹화된 단어들이 각각의 주제에 대해 선택되기 때문에 문서는 저 차원으로 위치하게 된다. 또한, 단어들은 독립적으로 발생하게 되고, 문서들은 무작위로 선택이 되기에, LDA는 문서의 문장/단어들의 맥락을 고려할 수 없게 된다. 단어의 차원과 문서의 차원의 두 차원이 있는 모델로 문서 벡터로

표현될 때 매우 희소한 표현을 하게 된다.

2. LDA2Vec

학습된 단어 임베딩은 단어의 특성이 벡터의 형식으로 학습되어 압축된 정보를 가지고 있지만, 그 자체로 해석하기에 어려움이 있다. 반면, LDA의 경우, 주제를 단어들의 조합으로 표현함으로서, 사용자에게 해석의 용이성을 준다. 상기 두 가지의 자연어 처리 방법을 동시에 분석에 적용한 LDA2vec은 2016년 Christopher Moody에 의해 2016년 제안된 방법으로, 전통적 토픽 모델링은 문맥적 의미를 반영할 수 있는 최신 분산적 단어 표현이 적용되지 않았던 반면, LDA2vec은 단어 임베딩을 통해, 단어 간의 관계를 유지하며 LDA의 토픽 모델링의 장점을 결합한 모델이다. 단어, 문서, 주제 벡터의 학습이 동일한 공간에서 동시에 이루어지며, 이는 문맥적 규칙 또한 보존하게 된다. [28] 그리고 해석 가능한 LDA 형식의 결과물을 제공하는 데 장점이 있다. LDA2vec의 손실 함수(1)는 다음과 같다.

$$\mathcal{L} = \mathcal{L}^d + \sum_{ij} \mathcal{L}_{ij}^{neg} \quad (1)$$

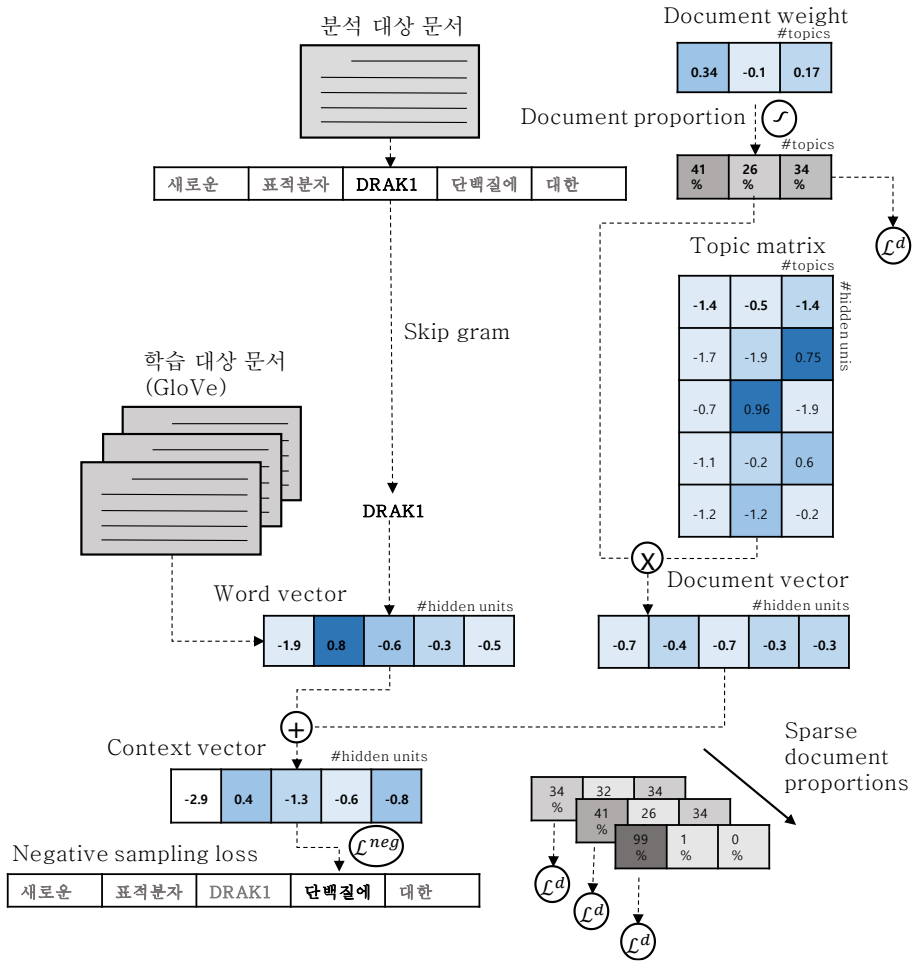
손실 함수는 Document Weight의 디리클레 우도함수와 Skipgram Negative Sampling Loss의 합으로 표현된다. 디리클레 우도함수(2)는 학습 진행 시, 희소성을 유지하는 역할을 하는데, 이는 Document Weight(p_{ij})가 모든 주제에 고루 분배되는 경향이 있기 때문이다.

$$\mathcal{L}^d = \lambda \sum_{jk} (\alpha - 1) log p_{jk} \quad (2)$$

Skipgram Negative Sampling Loss(3)는 컨텍스트 벡터 $\overrightarrow{(c_j)}$, 중심어 $\overrightarrow{(w_j)}$, 주변어 $\overrightarrow{(w_i)}$, 그리고 Negative Sampling Word$\overrightarrow{(w_l)}$으로 구성된다.

$$\mathcal{L}_{ij}^{neg} = log\sigma\left(\overrightarrow{c_j} \cdot \overrightarrow{w_i}\right) + \sum_{l=0}^{n} log\,\sigma\left(-\overrightarrow{c_j} \cdot \overrightarrow{w_l}\right) \quad (3)$$

단어들과 함께 컨텍스트 벡터들 또한 학습되게 되는데, 컨텍스트 벡터는



[그림 2-4] LDA2vec 모델 아키텍처 [28]

중심어와 문서 벡터의 합으로 주제에서의 중요도를 학습하게 된다. 문서 벡터는 기존 LDA와 같이 주제의 수가 하이퍼파라메터로 입력되어, 주제 벡터($\overrightarrow{t_k}$)와 Document Weight(p_{jk})(5)가 함께 학습된다.

$$\overrightarrow{c_j} = \overrightarrow{w_j} + \overrightarrow{d_j} \tag{4}$$

$$\overrightarrow{d_j} = p_{j0} \cdot \overrightarrow{t_0} + p_{j1} \cdot \overrightarrow{t_1} + \cdots + p_{jk} \cdot \overrightarrow{t_k} + \cdots + p_{jn} \cdot \overrightarrow{t_n} \tag{5}$$

LDA와 동일하게 하나의 문서는 여러 개의 주제를 내포할 수 있다는 가정을 동일 하게 적용하였다. Document Weight (p_{jk}) 들는 각각의

문서의 고유한 값이지만, 주제 벡터($\overline{t_k}$)들의 경우 문서 간의 공유로 문서 전체를 표현하게 된다. 이처럼 모든 벡터들이 하나의 공간에서 학습됨으로, 중심어가 주어졌을 때, 가장 가까운 주제 벡터를 신속하게 찾을 수 있는 장점이 있다.

3. Hybrid Topic Modeling

LDA2vec 이외에도 단어 임베딩과 LDA를 결합한 방식은 다양하게 연구되고 있다. Gaussian-LDA[29]의 경우, LDA2vec 보다 기존 LDA의 구조를 보존하면서, 문서는 단어 임베딩 값의 집합으로 가정하여, Gaussian-LDA를 구현한다. 단어 임베딩은 가우시안 분포를 중심으로 inverse Wishart 분포를 공분산으로 한 켤레 분포를 활용한다. 이후 깁스 샘플링을 통해 변수를 업데이트하고, 계산의 용이성을 위해 숄레스키 분해를 이용한 방법이다. 기존 LDA의 한계였던 OOV(Out Of Vocabulary)를 극복하고 성능 또한 향상되었다고 한다. 다른 연구는 LDA에서 얻어진 주제 벡터와 별도로 Word2vec을 이용해 문서를 학습시켜 문서가 주제와의 거리를 구하여 유사도를 분석하는 연구도 진행되었다. [30] 본 연구에서 LDA2vev을 선정한 이유는 하나의 공간에서 단어, 문서, 주제를 학습시킬 수 있고, 대량의 문서에서 학습된 단어 임베딩을 활용하여, 주제 상세 분석의 성능을 높일 방법으로 사료되어 선정하게 되었다.

제 3 장 데이터 분석

제 1 절 연구 제안서 데이터 수집 및 설명

1. 데이터 수집 및 분석 대상 선정 기준

국가과학기술지식정보서비스 (NTIS : National Science & Technology Information Service)를 통하여 2006년부터 2017년까지 정부에 의해 진행 또는 정부에 제안되어 진행된 이공계열 관련 데이터를 공유받았다. BT 연구과제 총 50,490건 중 바이오 신약개발 기술 분야 총 3,436건을 대상으로 진행하였으며, 최초 제안된 연구년을 기준으로 이후 진행 년에 대해서는 중복으로 판단하여 삭제를 진행하였다. 효과적인 추이 분석을 위해, 3개년을 하나의 기간으로 설정하여, 총 12개년을 4개의 구간으로 설정하여 분석을 진행하였다.

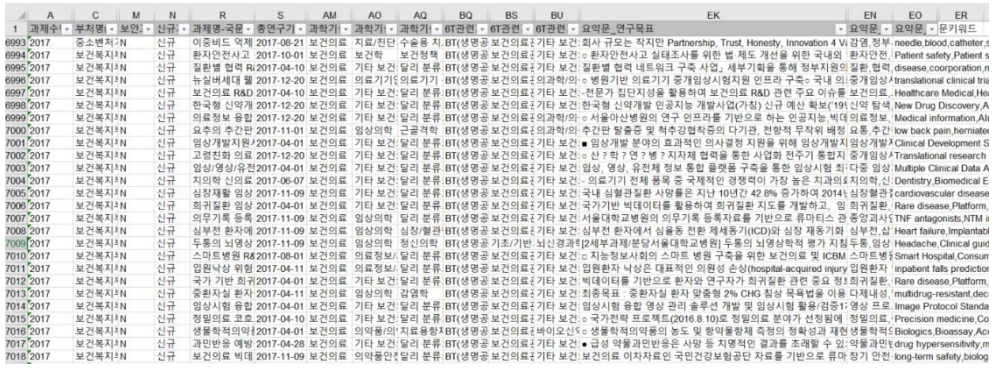
기간	연도	BT 데이터 – 대	바이오 신약 개발기술 – 소
기간 1	2006~2008	10,262 개	818 개
기간 2	2009~2011	10,824 개	715 개
기간 3	2012~2014	12,664 개	746 개
기간 4	2015~2017	16,740 개	1,157 개
총계	2006~2017	50,490 개	3,436 개

[표 3-1] 국가 연구과제 BT, 신약 분야별 데이터

2. 데이터 특성

대한민국 정부의 지원을 받아 진행되는 연구로, 국문 중심으로 데이터가 구성되어 있으며, 진행 정부 부서, 회계 구분, 연구 기간, 연구과제 구분, 상세 과제 구분 등 147개의 구분 열로 구성되어 있다. 하지만 기준별, 모든 정보를 보유하고 있지 않으며, 시점에 따라 부처의 생성과 소멸, 구분 방식이 상이하였다. 본 연구는 국문(자연어)처리를 통한 분석을 목표로 함으로, 개별 연구 목적을 확인할 수 있는

‘요약문_연구 목표’의 국문을 기본 데이터로 선정하였다. 또한, 연구가 진행된 첫해 년 도의 데이터만 선택하여, 본 연구 목표인 신약개발 방향에 집중하고자 하였다.



	A	C	M	N	R	S	AM	AO	AQ	BQ	BS	BU	EK	EN	EO	ER
1	과제수	부처명	보안	신규	과제명-국문	중연구기	과학기	과학기	과학기	6T관련	6T관련	6T관련	요약문_연구목표	요약문	요약문	문키워드
6993	2017	중소벤처기	N	신규	이중비드 억제	2017-08-21	보건의료	치료/진단	수술용 지	BT(생명공	보건의료	기타 보건	회사 규모는 작지만 Partnership, Trust, Honesty, Innovation 4 Vi	감염,정부	needle,blood,catheter,s	
6994	2017	보건복지부	N	신규	환자안전사고	2017-10-01	보건의료	보건학	보건정책	BT(생명공	보건의료	기타 보건	○ 환자안전사고 실태조사를 위한 법 제도 개선을 위한 국내외	환자안전	Patient safety,Patient s	
6995	2017	보건복지부	N	신규	질환별 협력 R	2017-04-10	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	질환별 협력 네트워크 구축 사업」 세부기획을 통해 정부지원의 질환 협력	disease,coorporation,n		
6996	2017	보건복지부	N	신규	뉴실버세대 웰	2017-12-20	보건의료	의료기기	의료기기	BT(생명공	보건의료	기타 보건	○ 병원기반 의료기기 중개임상시험지원 인프라 구축○ 국내 의	중개임상시	Translational clinical tris	
6997	2017	보건복지부	N	신규	보건의료 R&D	2017-04-10	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	·전문가 집단지성을 활용하여 보건의료 R&D 관련 주요 이슈를 보건의료	Healthcare Medical,Hea		
6998	2017	보건복지부	N	신규	한국형 신약개	2017-12-20	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	한국형 신약개발 인공지능 개발사업(가칭) 신규 예산 확보('19년 신약 탐색	New Drug Discovery,A		
6999	2017	보건복지부	N	신규	의료정보 융합	2017-12-20	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	의과학/의	○ 서울아산병원의 연구 인프라를 기반으로 하는 인공지능, 빅데	의료정보,	Medical information,Ali	
7000	2017	보건복지부	N	신규	요추의 추간판	2017-11-01	보건의료	임상의학	근골격학	BT(생명공	보건의료	의과학/의	추간판 탈출증 및 척추강협착증의 다기관, 전향적 무작위 배정 요통,추간	low back pain,herniate		
7001	2017	보건복지부	N	신규	임상개발지원	2017-04-01	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	■ 임상개발 분야의 효과적인 의사결정 지원을 위해 임상개발지	임상개발지	Clinical Development S	
7002	2017	보건복지부	N	신규	고령친화 의료	2017-12-20	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	○ 산 ? 학 ? 연 ? 병 ? 지자체 협력을 통한 사업화 전주기 통합지	중개임상시	Translational research	
7003	2017	보건복지부	N	신규	임상/영상/유전	2017-04-01	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	임상, 영상, 유전체 정보 통합 플랫폼 구축을 통한 임상시험 최	다중 임상	Multiple Clinical Data A	
7004	2017	보건복지부	N	신규	치의학 신의료	2017-06-07	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	- 의료기기 전체 품목 중 국제적인 경쟁력이 가장 높은 치과의료	치의학,신	Dentistry,Biomedical E	
7005	2017	보건복지부	N	신규	심장재활 임상	2017-11-09	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	국내 심혈관질환 사망률은 지난 10년간 42.8% 증가하여 2014년	심장혈관	cardiovascular disease	
7006	2017	보건복지부	N	신규	희귀질환 임상	2017-04-01	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	국가기반 빅데이터를 활용하여 희귀질환 지도를 개발하고, 임	희귀질환,	Rare disease,Platform,	
7007	2017	보건복지부	N	신규	의무기록 등록	2017-11-09	보건의료	임상의학	달리 분류	BT(생명공	보건의료	기타 보건	서울대학교병원의 의무기록 등록자료를 기반으로 류마티스 관	종양괴사인	TNF antagonists,NTM i	
7008	2017	보건복지부	N	신규	심부전 환자에	2017-11-09	보건의료	임상의학	심장/혈관	BT(생명공	보건의료	기타 보건	심부전 환자에서 심율동 전환 제세동기(ICD)와 심장 재동기화	심부전,삽	Heart failure,Implantabl	
7009	2017	보건복지부	N	신규	두통의 뇌영상	2017-11-09	보건의료	임상의학	정신의학	BT(생명공	기초/기반	뇌신경과학	2세부과제/분당서울대학교병원) 두통의 뇌영상학적 평가 지침두통,임상	Headache,Clinical guid		
7010	2017	보건복지부	N	신규	스마트병원 R&	2017-08-01	보건의료	의료정보/	달리 분류	BT(생명공	보건의료	기타 보건	□ 지능정보사회의 스마트 병원 구축을 위한 보건의료 및 ICBM	스마트병원	Smart Hospital,Consum	
7011	2017	보건복지부	N	신규	입원낙상 위험	2017-04-11	보건의료	의료정보/	달리 분류	BT(생명공	보건의료	기타 보건	입원환자 낙상은 대표적인 의원성 손상(hospital-acquired injury	입원환자	inpatient falls prediction	
7012	2017	보건복지부	N	신규	국가 기반 희귀	2017-04-01	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	빅데이터를 기반으로 환자와 연구자가 희귀질환 관련 중요 정	희귀질환,	Rare disease,Platform,	
7013	2017	보건복지부	N	신규	중환자실 환자	2017-04-11	보건의료	임상의학	감염학	BT(생명공	보건의료	기타 보건	최종목표 : 중환자실 환자 맞춤형 2% CHG 침상 목욕법을 이용 다제내성	multidrug-resistant,dec		
7014	2017	보건복지부	N	신규	임상시험 융합	2017-04-01	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	임상시험 융합 영상 관리 솔루션 개발 및 임상시험 활용/검증	영상 프로	Image Protocol Standa	
7015	2017	보건복지부	N	신규	정밀의료 프로	2017-04-10	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	○ 국가전략 프로젝트(2016.8.10로 정밀의료 분야가 선정됨에	정밀의료,	Precision medicine,Co	
7016	2017	보건복지부	N	신규	생물학적의약	2017-04-01	보건의료	의약품/의	치료용항체	BT(생명공	보건의료	바이오신약	○ 생물학적의약품의 농도 및 항약물항체 측정의 정확성과 재현	생물학적의	Biologics,Bioassay,Acc	
7017	2017	보건복지부	N	신규	과민반응 예방	2017-04-28	보건의료	기타 보건	달리 분류	BT(생명공	보건의료	기타 보건	● 급성 약물과민반응은 사망 등 치명적인 결과를 초래할 수 있	약물과민반	drug hypersensitivity,m	
7018	2017	보건복지부	N	신규	보건의료 빅데	2017-11-09	보건의료	의약품안전	달리 분류	BT(생명공	보건의료	기타 보건	보건의료 이차자료인 국민건강보험공단 자료를 기반으로 류마	장기 안전	long-term safety,biolog	

[그림 3-1] 국가 연구과제 NTIS 공유 기준 데이터

제 2 절 기본 전처리 과정

1. 기본 전처리

1.1. 문서 전처리

NTIS에서 공유된 데이터의 포맷은 엑셀로 일부 정보의 누락과 보안 관련된 과제들은 내용을 미포함하고 있었다. 상기 1절의 데이터 선정 기준에 따른 대분류 6대 중점 과제 중 BT를 선정하였고, 대, 중, 소분류 중, 상세 분류 항목인 소분류에서 ‘바이오 신약개발기술’로 선정렬한 후, 정보가 비어있는 행은 삭제를 진행하였다. 또한, 이공계와 연관성이 낮은 관련 부처(산림청, 고용노동부 등)의 연구는 삭제를 진행하여, 데이터가 존재하고 분석에 필요한 분야에 집중하고자 하였다.

1.2. 분석 단어 선정

1.1에서 1차 문서 정제 진행 이후, 리스트 형태로 데이터 포맷을 변경하였고, Python의 자연어 처리 패키지인 nltk를 활용하여, 토큰화(Tokenization)를 진행하였다. 이후, 한국어 자연어 처리 패키지 ckonlpy를 활용하여, 품사(POS-Tagging)를 부여하였다. 명사, 조사,

형용사, 영어 등으로 분리된 토큰 중, 문서 내의 주제를 대표하는 의미를 표현 할 수 있는 명사와 영어를 선정하여 분석 하였다.

1.3. 불용어 처리

불용어 처리는 분석 단어 품사를 선정한 후, 최적의 주제의 수를 찾기 위한 Coherence Score의 결괏값을 적용한 LDA의 결과를 관찰하며, 불용어에 대한 처리를 진행하였다. 의미가 분석의 방향에 영향을 낮게 미칠 것으로 예상되는 공통적인 표현을 삭제 우선순위로 진행하였다. '연구', '분석', '도출하고자', '1차년도' 등이 이에 포함된다. 또한, 효율적인 불용어 처리를 위해 Skip-gram을 활용하여, 서로의 영향을 많이 주는 단어의 빈도를 분석하여, 불용어 처리를 진행하였다. LDA2vec의 경우 GloVe 학습 데이터를 기준으로 불용어 처리를 진행하였으며, 약 15만 건의 데이터에서 5개 미만 출현 빈도 한글, 영어 단어, 3개의 소문자로 이루어진 영문과 상위 출현 빈도 단어 중 LDA진행 시 분석 방향에 영향이 낮은 단어들에 대한 삭제를 진행한 데이터로 15만 건 학습 문서 및 선정 기간의 문서를 정제하였다.

2. 전문 언어를 위한 전처리

분석 목표 데이터가 전문적인 내용을 포함할수록, 형태소 분석기를 통한 정보 추출을 하기 어렵다. 일반 형태소 분석기는 포함 단어의 한계가 있어, 데이터 특성에 맞는 단어를 추가해주는 것이 효율적이다. 이에 바이오 분야에 특화된 언어의 처리를 위하여, n개의 토큰 (형태소 분석기가 문장을 단어로 나누는 기본 단위)으로 묶어 함께 처리하는 n-gram보다 상호정보량을 이용하여 대상 데이터에 맞는 단어를 찾을 수 있는 점별상호정보량, PMI (pointwise mutual information)(1)을 활용하여, 분석기를 갱신하고자 하였다. 상호정보량(Mutual Information)은 두 확률 변수의 관계를 나타내는 정보량이다. 이는 두 개의 다른 토큰(A, B)이 함께 등장하는 정보량을 바탕으로 토큰을 묶어 하나의 토큰을 생성할 수 있다. 각 토큰의 등장 확률(p(A), p(B))과 동시 등장 확률 $p(A \cap B)$ 을 통한 계산과 전체 데이터의 토큰 수(n)와 단어 등장 횟수(a, b)로 계산하는 두 가지 방식으로 PMI는 다음과 같이 정의된다.[31] 높은 PMI일수록 동시 발생 확률이 높다.

$$PMI(A,B) = log\frac{p(A\cap B)}{p(A)p(B)} = log\frac{nz}{ab} \quad (1)$$

이때, 전체 데이터의 토큰 수(n)에 따라 PMI의 값이 달라진다. 이에 다른 데이터에서의 PMI 값과 비교가 어려워 정규화가 필요하다. PMI 값을 [−1, 1]의 범위로 정규화하는 방법은 다음과 같은 식(2)을 사용하였다.

$$NPMI(A,B) = \frac{PMI(A,B)}{-\log p(A\cap B)} = log\frac{log\frac{nz}{ab}}{log\frac{n}{z}} \quad (2)$$

다음의 NPMI(Normalized pointwise mutual information)식은 인접한 두 개의 토큰을 기준으로 계산 하여, 3개의 토큰을 활용한 연어 추출에 사용된 식이다. 다변수 상호정보량 (Multivariate Mutual Information)(3)을 활용하여 계산하였다. [32]

$$NPMI(A,B,C) = \frac{PMI(A,B,C)}{-2\log p(A\cap B\cap C)} = log\frac{log\frac{n^2z}{abc}}{2log\frac{n}{z}} \quad (3)$$

바이오 분야는 화학 명칭의 조합, 단백질 명들의 조합, 연구 분야의 전문 용어들의 조합으로 신규 분야 확장이 이루어져, 합성 명사의 분석이 중요하다. 이에 본 연구에서는 PMI 기법을 활용하여 두 개의 토큰, 세 개의 토큰을 NPMI로 묶어주어 연어를 생성하였으며, 이를 형태소 분석기 사전에 추가하여, 바이오 데이터에 전문화된 토큰을 재생성하였다. 하기 표 3−2와 표 3−3는 NPMI를 적용하여 생성된 단어들의 예시이다. T cell, 만성 어깨 파열, 자기공명영상 등 바이오 헬스케어 분야에 특화된 단어들을 확인할 수 있었다.

토큰 1	토큰 2	명사	NPMI
폴리	카프로락톤	폴리카프로락톤	0.961
딥	러닝	딥러닝	0.952
E.	coli	E.coli	0.948
B 형	간염	B 형간염	0.899
T	cell	Tcell	0.63

[표 3−2] NPMI 2단어 결합 예시

토큰 1	토큰 2	토큰 3	명사	NPMI
만성	어깨	파열	만성어깨파열	0.752
자기	공명	영상	자기공명영상	0.708
제	2 형	당뇨병	제 2 형당뇨병	0.693
급성	심부전	세포	급성심부전세포	0.56

[표 3-3] NPMI 3단어 결합 예시

3. 주제 탐색 강화 (TF-IDF)

TF-IDF(Term Frequency -Inverse Document Frequency)는 단어 빈도와 역문서 빈도의 곱으로, 문서에서 특정 단어의 중요도를 통계적 수치로 확인 할 수 있는 방법이다. TF-IDF는 가중치로 정보 검색, 텍스트 마이닝에 사용된다.

$$tfidf(t,d,D) = \left[0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d):w \in d\}}\right] \times log\frac{|D|}{|\{d \in D: t \in d\}|} \quad (1)$$

본 연구에서는 LDA 분석 전, 문서와 단어의 행렬의 조합에 TF-IDF를 적용하여, 단어의 중요도를 적용한 후, LDA 분석을 진행하였다. 이는 LDA 분석의 주제 탐색의 성능을 높이기 위한 방법으로 사용하였다.

제 3 절 분석 파이프라인

1. LDA

Python 3.7을 기반으로 Anaconda 4.7.11의 Spyder를 사용하였다. 3장 1절에서 언급한 3개년을 하나의 기간으로 설정하고, 바이오 신약개발 기술 분야를 분석 문자 데이터로 설정하였다. 총 4개의 코퍼스가 생성되었고, 각 각의 코퍼스에 대해 토픽 모델링(LDA) 분석을 진행하였다. 프로그래밍 순서는 1) Pandas로 data frame 변경 2) nltk 3.4.3 토큰화 진행 3) cknolpy 0.0.64로 한글 품사 표기 진행 (NPMI 에서 생성된 코퍼스 특화 단어 명사로 추가) 4) 불용어 처리 5) 명사와 영문 이외 품사 삭제 6) tf-idf으로 LDA 입력 코퍼스 정제 7) gensim

3.7.3 이용 LDA 분석 진행하였다. 주제의 수의 경우, Coherence Score를 축으로 주제의 수와 그래프 화하여, Coherence Score가 높은 주제 수를 선정하여 분석 진행하였다.

2. LDA2Vec

Python 3.5와 Tensorflow – gpu 1.12.0을 환경으로 Pycharm을 사용하여 학습을 진행하였다. GPU는 GeForce RTX 2080를 사용하였다. 1) GloVe 학습 데이터는 총 151,341건으로 학습 데이터를 최대화하기 위해 연구 대상인 '요약문_연구목표'(50,478건) 이외에 동일한 연구에 포함되어 연구에 대한 상세 내용과 기대효과를 기술한 '요약문_기대효과' (50,401건), '요약문_연구내용'(50,460건)을 활용하였다. 총 151,341건의 학습 데이터를 LDA 진행 시 진행한 동일한 전처리 기법으로, 불용어 제거 및 명사와 영문만 포함된 코퍼스 생성하였다. 2) 이후, GloVe로 총 151,340건의 문건, 93,060개 토큰을 300차원으로 1,000 epoch 학습시켰다. 이후 LDA2Vec 알고리즘에 GloVe로 학습된 임베딩 데이터와 3개년으로 묶인 4개의 코퍼스를 각각 학습 시켜 기간별 분석을 진행하였다. 분석 과정에서의 하이퍼파라메터인 학습 epoch와 lambda, alpha, learning rate, switch_loss_epoch는 한글 코퍼스에 맞게 최적화를 진행하였다.

종류	학습 데이터	기간1	기간2	기간3	기간4
문서	151,340	818	715	746	1,157
단어수	93,060	4,800	4,249	4,753	6,414

[표 3-4] 학습 데이터 및 기간별 단어 수

제 4 장 구현과 결과 분석

3장까지의 전처리 과정을 통하여, 합성명사 및 코퍼스에 특화된 제약 산업 관련 단어들을 찾아 주요 단어들이 토픽 모델링 과정에서 포함되게 하여 분석 결과의 이해도를 높이고자 하였다. 이후 두 가지 토픽 모델링 기법(LDA, LDA2vec)의 구현을 완료하였다. 본 장에서는 기간별 주제의 변화 추이를 분석하고 모델링 기법 간의 특징에 관해 서술함과 동시에, LDA2vec 방법을 활용하여 상세 주제 변화 탐지와 추후 주제에 대한 예측 가능성을 보고자 한다.

제 1 절 NPMI 적용 결과

NPMI 적용 결과는 기간 4에 한정하여 성능 확인을 진행하였다. 동일한 조건 (불용어, LDA 하이퍼파라메터)에서 ckonlpy의 명사 사전에 NPMI로 추출된 합성 명사의 추가 여부를 조건으로 설정하여 진행하였다. 표 4-1과 표 4-2는 신규 합성 명사의 적용 전과 후의 결과이다. 동일한 조건에서 LDA를 실행하였기에, 동일 분포의 단어가 표기되어야 하지만, 문서에 특화된 단어의 추가는 암 줄기세포, CD8(세포표면항원무리 8, 당단백질), 바이오시밀러, 조골세포, 류머티즘 관절염과 같은 코퍼스에 특화된 단어들이 표 4-2에 표현되었다. 기존 줄기세포, 바이오, 세포, 관절염 등에 기반한 분석보다 한 단계 나아간 상세 정보를 제공함을 관찰 할 수 있다.

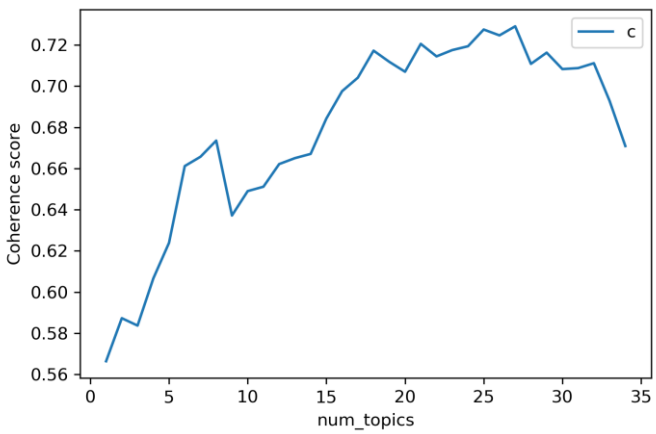
단어	단어	단어	단어	단어
줄기세포	대장암	대사성	플랫폼	중소기업
스케일	HDAC	RANKL	유전체	소프트웨어
CAR	regulatory	아데노바이러스	repebody	assembly
유전학	Data	자가면역	oncogenic	구조체
피부염	아토피	페스트균	CMC	MERS
루푸스	림프구	albicans	RNA	CoV
당뇨병	PET	영장류	안구건조증	에너지
발기부전	압타머	VEGF	Prx	췌장암
패혈증	소화기	섬유아세포	지적재산권	섬유화

[표 4-1] NPMI 적용 전, 기간 4

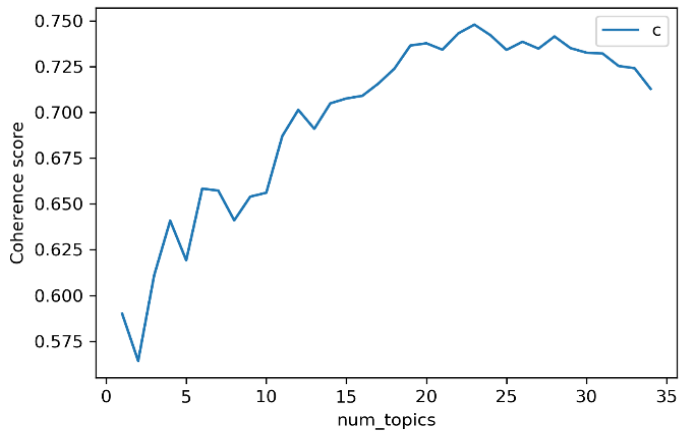
단어	단어	단어	단어	단어
지적재산권	항바이러스제	항체절편	알고리즘	분자표적
Hit	감염병	CD8	당대사	면역세포치료제
간세포암	관절염	바이오의약품	제약회사	류마티스
안구건조증	BBB	류마티스관절염	파킨슨	한국인
바이러스	중간엽줄기세포	항체치료제	면역원성	화장품
면역억제제	융합단백질	바이오이미징	백혈병	호흡기
MERS	망막병증	CoV	조절물질	항체의약품
glioblastoma	가이드라인	SFTSV	HIV	혈관내피세포
GPCR	돌연변이	콜라겐	이중항체	항암항체

[표 4-2] NPMI 적용 후, 기간 4

또한, 주제의 수를 산정하는데 활용한 Coherence Score도 NPMI 적용 전 그림 4-1보다 상승한, NPMI 적용 후 그림 4-2에서 최대치 값이 높음을 확인 할 수 있다. 동일한 주제 수에서 0.72에서 0.75까지 증가하였으며, 이는 문서 간의 의미적 연관성이 높아지며 군집화되었음을 나타낸다.



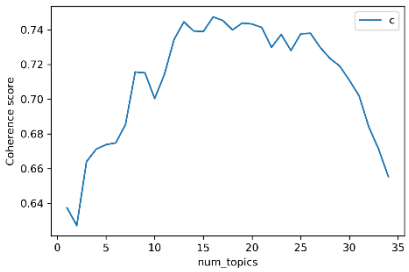
[그림 4-1] NPMI 미적용 Coherence Score



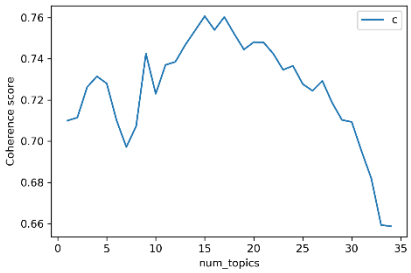
[그림 4-2] NPMI 적용 Coherence Score

제 2 절 LDA 분석 결과

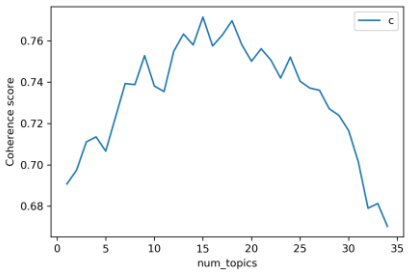
LDA의 하이퍼파라메터인 주제의 수는 2010년 D. Newman에 의해 제시된 방법을 참고하였다. 주제로 뽑힌 상위 단어들의 유사도가 의미론적으로도 연관성이 있다 가정한 Coherence Score[33]를 계산하여 고점의 Coherence Score 주변의 주제의 수를 참고하고자 하였다. 본 분석에서는 참고 코퍼스가 없이도 Coherence Score를 유도할 수 있는 M. Röder[34]의 Coherence Score를 활용하였다. (genism package) 그래프로 표현된 주제의 수의 변화에 따른 Coherence 값은 그림 4-3과 같다. 최고점의 Coherence Score를 기준으로 주변의 주제 수로 변경하며, 한 주제 다양한 주제명이 포함되지 않도록 하는 최적의 주제의 수를 찾고자 하였다. 또한 Coherence Score가 0.7 이상으로 최적화를 진행 하여, 의미를 보존하며 군집화하고자 하였다.[35] 본 방법으로 기간별 주제 수를 설정하였고, 결과는 표 4-3과 같다. Coherence Score의 최적의 주제 수로 LDA에 적용한 이후 기간별 총 문헌에서의 주제별 점유율을 분석하였다. LDA에서 하나의 문헌은 여러 개의 주제로 이루어진다는 가정을 한다. 이때, 가장 비중이 큰 주제를 각 문헌의 대표 주제로 판단하고, 이를 각각의 문헌에 적용하여, 전체 문헌에서 주제들이 가지는 비중을 산출하였다. 이를 표 4-4, 4-5, 4-6, 4-7에 LDA의 결과 단어들과 함께 표기하였다. 이는 그림 4-4의 기간별 주제의 비중 표현에 활용된다.



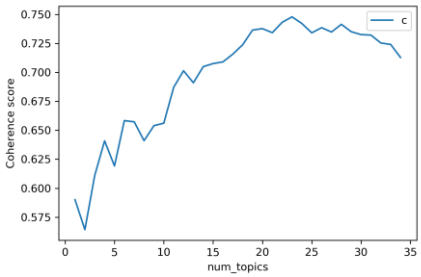
<2006~2008년 Cohernece Score>



<2009~2011년 Cohernece Score>



<2012~2014년 Cohernece Score>



<2015~2017년 Cohernece Score>

[그림 4-3] 기간 별, Coherence Score

구분된 주제에 대한 명칭은 각각의 단어의 주제에 대한 기여도를 기반으로 선정하였으며, 일부 주제의 경우 두 가지 주제명으로 해석이 가능할 수 있는데, 이 경우는 상위 기여도 단어에 우선순위를 부여하고, 이후 표현된 연관 단어와의 해석의 용이함을 기준으로 명칭을 설정하였다.

기간	연도	선택 주제 수	Coherence Score
기간 1	2006~2008	12	0.72
기간 2	2009~2011	13	0.75
기간 3	2012~2014	13	0.76
기간 4	2015~2017	16	0.72

[표 4-3] 기간 별, 선정 주제 수와 Coherence Score

비중	기여도	단어	기여도	단어	기여도	단어	기여도	단어	기여도	단어
6.00%	0.005	대장암	0.005	신약개발	0.005	파킨슨병	0.004	미생물	0.004	유산균
	0.004	동맥경화	0.003	표적단백질	0.003	이미징	0.003	경구투여	0.003	중후군
13.50%	0.012	펩타이드	0.009	천연물	0.009	유방암	0.007	항암제	0.005	세포사멸
	0.005	단클론	0.004	제약기업	0.003	prodrug	0.003	아미노산	0.003	고분자
7.00%	0.004	FDA	0.004	Romo1	0.004	hypoxia	0.004	항체생산	0.004	악성화
	0.003	대사물질	0.003	IND	0.003	융합단백질	0.003	형질전환	0.003	항암제
10.40%	0.019	골다공증	0.008	라이브러리	0.006	항생제	0.005	바이러스	0.005	조골세포
	0.005	전문인력	0.004	제품화	0.004	골형성	0.004	당단백질	0.004	peptide
6.80%	0.009	경구용	0.007	맞춤형	0.006	PGE	0.005	IND	0.004	HBV
	0.004	Foxp3	0.004	일본뇌염	0.004	Prostaglandin	0.004	hGH	0.003	글리벡
8.20%	0.007	바이러스	0.007	조류인플루엔자	0.005	인터페론	0.005	돌연변이	0.005	인플루엔자
	0.004	뇌졸중	0.004	HER2	0.004	siRNA	0.004	대유행	0.003	약효지속성
9.60%	0.011	IND	0.005	신경병증	0.005	면역원성	0.005	당뇨병	0.005	안구건조증
	0.004	Docetaxel	0.004	길항제	0.004	인력양성	0.004	신약개발	0.004	바이러스
9.50%	0.008	바이오마커	0.006	분자기전	0.006	암세포	0.005	박테리아	0.005	리간드
	0.004	DDS	0.004	독성평가	0.004	방사성의약품	0.004	입체구조	0.004	Wnt
7.70%	0.006	위장관	0.005	조절물질	0.005	HTS	0.004	감염성질환	0.004	생체기능
	0.004	신약설계	0.004	FAF1	0.004	암관련	0.004	허혈성질환	0.003	초고속
7.60%	0.008	비만치료제	0.008	난치성	0.008	MCH	0.006	Melanin	0.005	주사제
	0.004	ECS	0.003	단백질체학	0.003	비만세포	0.003	선도물질개발	0.003	알츠하이머
6.50%	0.007	HCV	0.006	NF-κB	0.005	DNA	0.005	말라리아	0.004	감염성
	0.004	암치료	0.004	항진균	0.004	임상연구	0.003	항진균제	0.003	약물수송체
7.10%	0.007	류마티스	0.005	관절염	0.004	항비만	0.004	바이오의약품	0.003	대량생산
	0.003	약물상호작용	0.003	줄기세포	0.003	제품화	0.003	개량신약	0.003	치료백신

[표 4-4] 기간1, 주제별 문서 비중과 내용

상기의 기준으로 LDA의 주제명을 설정하였고, 관련이 높은 주제와는 기간별 연결을 진행하였다. 그림 4-4는 기간별 주제명 분석을 통해 동일 주제를 연결한 도표이다. 주제의 생성과 소멸, 그리고 기간별 주제의 비중의 변화가 표현되어 있다. 이를 통하여 신약 개발 분야의 연구 방향을 분석 할 수 있다.

비중	기여도	단어	기여도	단어	기여도	단어	기여도	단어	기여도	단어
5.30%	0.014	골다공증	0.007	조골세포	0.006	항당뇨	0.005	antibody	0.005	G-CSF
	0.005	파골세포	0.004	바이러스	0.004	liver	0.003	혈소판	0.003	PPAR
14.9%	0.014	항암제	0.01	유방암	0.008	신약후보물질	0.008	위장관	0.007	나노젤
	0.006	항산화	0.006	치료백신	0.006	IL-10	0.005	전립선암	0.005	단백체
9.00%	0.017	바이러스	0.016	인플루엔자	0.008	바이오의약품	0.008	신약개발	0.008	뇌졸중
	0.007	파킨슨병	0.006	KFDA	0.006	상용화	0.005	약물전달체	0.005	곤충세포
9.6%	0.009	나노입자	0.005	마이셀	0.005	암세포	0.005	BAF	0.005	췌장암
	0.005	질환동물	0.005	간질환	0.005	노로바이러스	0.005	허셉틴	0.004	제품화하고자
7.60%	0.007	방광암	0.006	HSP27	0.005	항암활성	0.005	유전자치료제	0.004	대유행
	0.004	제품화	0.004	돌연변이체	0.004	항비만효과	0.004	항체의약품	0.003	기반기술구축
7.60%	0.015	천연물	0.008	천연물신약	0.007	감염성	0.007	아토피	0.006	바이러스
	0.006	퇴행성	0.005	플라보노이드	0.005	골형성	0.005	신약후보	0.004	동맥경화증
5.80%	0.008	PAUF	0.006	이중표적항체	0.005	항체치료제	0.004	면역보조제	0.004	섬유화
	0.004	내성균	0.004	TNF-α	0.003	균막형성능	0.003	로타바이러스	0.003	국내임상
6.30%	0.009	플랫폼	0.007	동맥경화	0.007	HDL	0.005	콜레스테롤	0.005	항바이러스
	0.005	천연물	0.004	고지혈증	0.004	apolipoprotein	0.004	물성분석	0.004	심바스타틴
6.30%	0.011	항염증	0.01	화장품	0.006	GMP	0.005	TLR	0.005	해양천연물
	0.005	심혈관계	0.005	칼페인	0.005	GPR	0.005	간섬유화	0.005	항비만
4.50%	0.007	VEGF	0.007	인간항체	0.006	약물전달시스템	0.005	PGE	0.005	항암화학요법
	0.004	피부질환	0.004	제생인자	0.003	Prostaglandin	0.003	약물방출	0.003	생체적합성/생
7.30%	0.011	DDS	0.01	비만치료제	0.009	라이브러리	0.009	제품화	0.007	암환자
	0.007	Melanin	0.007	MCH	0.007	신약개발	0.006	신의약	0.005	항생물질
8.70%	0.012	줄기세포	0.01	당뇨병	0.009	폐혈증	0.008	TRAIL	0.007	암세포
	0.007	관절염	0.005	cGMP	0.005	PI3K	0.005	HBV	0.004	당뇨치료제
7.00%	0.01	바이오시밀러	0.009	항암제	0.008	인터페론	0.008	미토콘드리아	0.007	표준화
	0.006	대사증후군	0.005	HCV	0.005	11β-HSD1	0.005	protease	0.005	암세포

[표 4-5] 기간2, 주제별 문서 비중과 내용

의약품은 시대 별 4가지의 형태로 진화 하였는데[36], 초기의 전통 의약품, 두 번째로 합성 본 연구의 데이터는 세 번째 바이오 의약품의 시대와 유전자 기반 신약의 시대에 포함되어 있다 할 수 있다. 이에, 바이오 의약품에 해당하는 항체, 백신, 단백질을 이용한 연구가

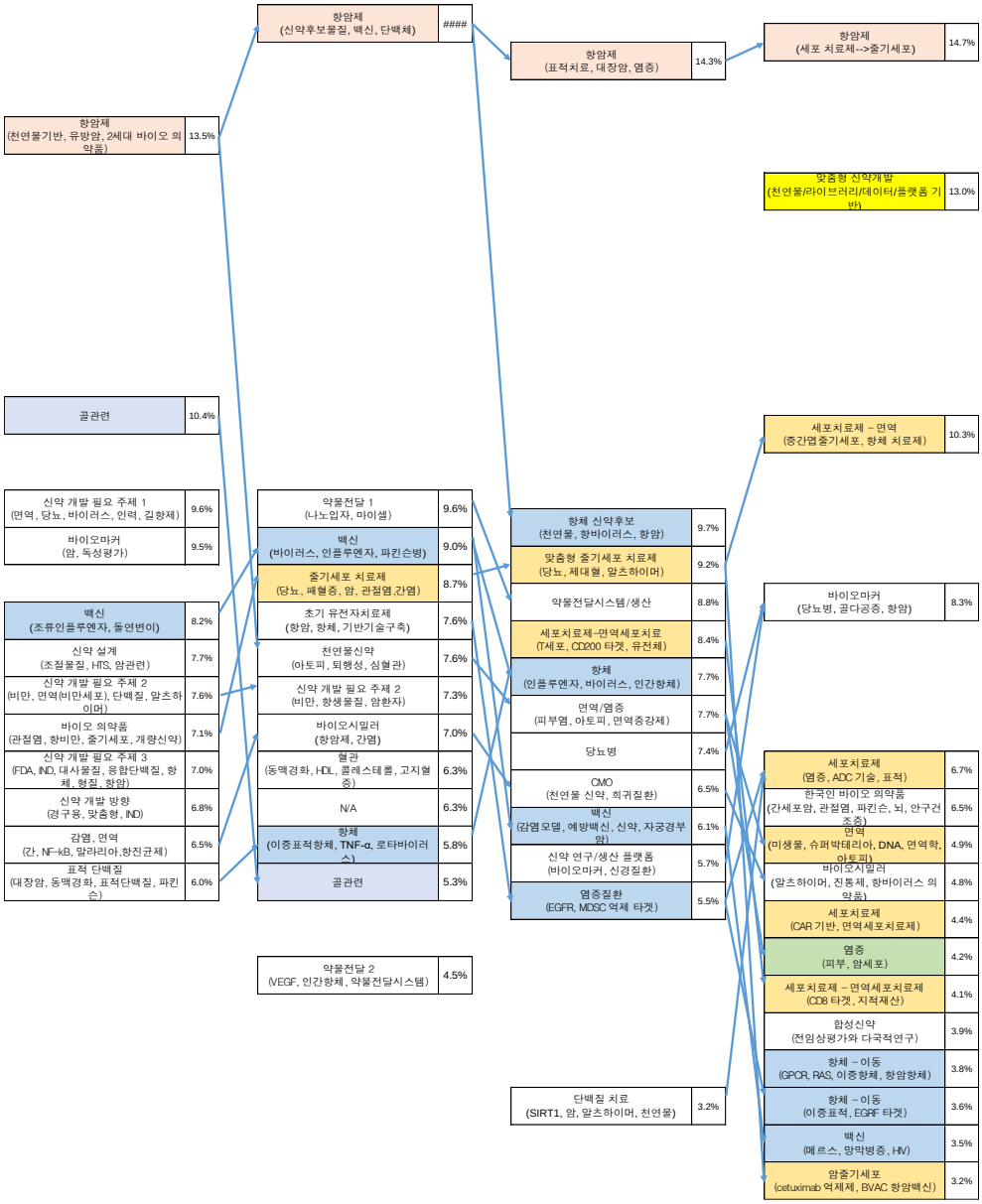
비중	기여도	단어	기여도	단어	기여도	단어	기여도	단어	기여도	단어
9.20%	0.015	줄기세포	0.006	당뇨병	0.006	항암제	0.005	맞춤형	0.005	제대혈
	0.004	상피세포	0.004	알츠하이머병	0.004	약물동태	0.004	배양액	0.004	POC
6.10%	0.008	바이오이미징	0.008	감염모델	0.005	서비스	0.004	Telmisartan	0.004	TCTP
	0.004	예방백신	0.004	신약개발	0.003	자궁경부암	0.003	EMT	0.003	알레르기
6.50%	0.01	Wnt	0.008	CMO	0.007	신호전달계	0.005	천연물신약	0.004	희귀질환
	0.004	제약회사	0.004	단일클론항체	0.003	ALS	0.003	SAD	0.003	cilia
9.70%	0.015	바이러스	0.008	표준화	0.007	천연물	0.006	신약후보물질	0.005	HBV
	0.005	항바이러스	0.005	항암제	0.005	효능검증	0.005	DNA	0.004	암세포
8.80%	0.011	압타머	0.01	대사성	0.008	나노입자	0.007	방사성의약품	0.007	암세포
	0.006	아미노산	0.006	독감바이러스	0.005	대량생산	0.005	HER2	0.004	RNA
7.40%	0.011	인슐린	0.01	당뇨병	0.006	제2형	0.005	접합체	0.005	에스트로겐
	0.005	폐혈증	0.004	SOP	0.004	GPR40	0.004	IRB	0.004	Treg
7.70%	0.014	인플루엔자	0.005	바이러스	0.005	GPR	0.005	인간항체	0.005	항체의약품
	0.005	국제공동연구	0.005	항체치료제	0.004	허혈성	0.004	프로테아좀	0.004	항생제
7.70%	0.007	면역증강제	0.007	아토피	0.006	MRI	0.005	의료기기	0.005	유방암
	0.004	피부염	0.004	림프종	0.003	WHO	0.003	제품화	0.003	복합제
14.30%	0.009	항암제	0.007	GMP	0.006	당사슬	0.006	류마티스	0.006	인산화
	0.006	표적치료	0.005	대장암	0.004	관절염	0.004	FlaB	0.004	시제품
5.70%	0.011	플랫폼	0.009	컨설팅	0.007	대사질환	0.005	파이프라인	0.004	뇌질환
	0.004	신경병증	0.004	바이오마커	0.004	대사체	0.003	슈퍼박테리아	0.003	뇌신경계
8.40%	0.008	태상포진	0.007	T세포	0.007	유전체	0.005	바이오의약품	0.004	대식세포
	0.004	면역원성	0.004	CD200	0.004	글립틴	0.003	iron	0.003	IND신청
5.50%	0.006	EGFR	0.005	자가면역질환	0.005	류마티스관절염	0.004	세포소기관	0.004	MDSC
	0.004	병태생리학적	0.004	자궁내막암	0.004	염증성	0.004	omics	0.004	장질환
3.20%	0.007	알츠하이머	0.006	SCF	0.005	결합단백질	0.004	약제내성	0.004	ASC
	0.004	천연물	0.004	암세포주	0.003	소프트웨어	0.003	SIRT1	0.003	phlorogluc

[표 4-6] 기간3, 주제별 문서 비중과 내용

많았으며, 바이오 의약품의 1세대 (체내 부족 단백질 대응, 인슐린, 백신 관련 주제), 2세대 (인체 미 존재 단백질 생성, 동물세포 활용, 융합 단백질, 항암, 면역 관련)부터, 3세대인 세포치료제, 유전자 치료제(면역세포 치료제, 줄기세포 치료제 관련)의 연구 방향도 탐색 할 수 있었다.

비중	기여도	단어	기여도	단어	기여도	단어	기여도	단어	기여도	단어
14.70%	0.022	항암제	0.019	줄기세포	0.013	암세포	0.011	미세환경	0.01	대량생산
	0.009	영장류	0.009	항염증	0.006	전립선암	0.005	항생제	0.005	에소좀
3.20%	0.01	암줄기세포	0.008	시제품	0.006	자가면역질환	0.005	항암활성	0.005	cetuximab
	0.005	Peptide	0.005	BVAC	0.005	이차대사산물	0.004	인간화	0.003	암항원
6.70%	0.014	염증성	0.009	제조공정	0.008	신경세포	0.008	GLP	0.008	ADC
	0.007	표적형	0.007	공정개발	0.007	세포치료제의	0.006	대사질환	0.006	장질환
4.10%	0.006	지적재산권	0.005	항바이러스제	0.005	항체절편	0.005	알고리즘	0.005	분자표적
	0.004	Hit	0.004	감염병	0.004	CD8	0.004	당대사	0.004	면역세포치료제
3.90%	0.008	박테리아	0.007	ALK	0.006	유방암	0.006	식약처	0.006	다국적
	0.006	세포독성	0.006	합성신약	0.005	혈관신생	0.004	HER2	0.004	극대화할
4.80%	0.01	바이오시밀러	0.006	제품화	0.006	알츠하이머	0.005	당사슬	0.005	한약재
	0.005	RANKL	0.005	항바이러스	0.005	개량신약	0.005	난소암	0.005	진통제
3.60%	0.009	EGFR	0.008	건강기능식품	0.006	PEG	0.006	나노입자	0.005	이중표적
	0.005	IPF	0.005	NOX	0.004	조골세포	0.004	다제내성균	0.004	대장암
8.30%	0.015	인슐린	0.012	RNA	0.011	미토콘드리아	0.01	골다공증	0.009	항암치료
	0.009	바이오마커	0.009	폐혈증	0.008	당뇨병	0.007	뇌종양	0.006	설치류
4.90%	0.011	미생물	0.008	DNA	0.008	뇌졸중	0.007	면역학	0.006	Autophagy
	0.006	슈퍼박테리아	0.005	디스플레이	0.005	면역반응	0.005	아토피피부염	0.004	비만세포
4.40%	0.012	PD-L1	0.008	CAR	0.006	세포치료제	0.005	인간항체	0.004	면역치료제
	0.004	면역증강	0.004	바이오마커	0.004	허혈성	0.004	융합연구	0.004	immune
13.00%	0.015	맞춤형	0.015	천연물	0.014	라이브러리	0.011	신약개발	0.01	인플루엔자
	0.009	인터페론	0.008	바이러스	0.008	대장암	0.007	데이터	0.007	플랫폼
6.50%	0.013	간세포암	0.012	관절염	0.01	바이오의약품	0.008	제약회사	0.007	류마티스
	0.007	안구건조증	0.006	BBB	0.006	류마티스관절염	0.006	파킨슨	0.006	한국인
10.30%	0.016	바이러스	0.012	중간엽줄기세포	0.01	항체치료제	0.009	면역원성	0.009	화장품
	0.009	면역억제제	0.009	융합단백질	0.008	바이오이미징	0.008	백혈병	0.008	호흡기
3.50%	0.008	MERS	0.007	망막병증	0.007	CoV	0.006	조절물질	0.004	항체의약품
	0.004	glioblastoma	0.004	가이드라인	0.003	SFTSV	0.003	HIV	0.002	혈관내피세포
3.80%	0.014	GPCR	0.012	돌연변이	0.006	콜라겐	0.006	이중항체	0.005	항암항체
	0.005	간질환	0.004	RAS	0.004	합병증	0.003	변이체	0.003	자가면역
4.20%	0.019	피부염	0.014	아토피	0.009	중소기업	0.007	sili	0.005	소프트웨어
	0.005	brain	0.005	케모카인	0.004	PDK	0.004	암세포주	0.003	포도당

[표 4-7] 기간4, 주제별 문서 비중과 내용



[그림 4-4] 기간 별 문서의 비중과 동일 주제의 변화

기간별 특징은 기간 1은 신약 개발의 주제가 혼합되어 있는 경향이 높았다. 이에 세부적인 물질이나 연구 분야를 구분하기 어려웠다. 하지만, HTS(High Throughput Screening)나 IND(Investigational New Drug)의 주제 단어로 보았을 때, 신약 후보군에 대한 탐색이

주요했던 시기라 할 수 있다. 기간 2는 신약 개발의 주요 주제들을 확인할 수 있다. 줄기세포 치료제, 유전자 치료제, 천연물신약, 바이오시밀러, 항체 등의 주제명이 도출되었다. 나아가 주제의 세부 내용도 확인이 가능한데, 항체의 경우는 이중 표적, TNF-α (종양 괴사 인자 알파) 등의 단어로 기술의 추이나 표적 하는 단백질 명도 확인 가능하였다. 최근 국내에서 바이오산업을 이끄는 바이오시밀러는 항암제, 간염과 함께 주제를 형성하였다. 기간 3에서는 CMO(Contract Manufacturing Organization)와 신약 연구/생산 플랫폼 등 사업화에 대한 내용이 출현하였다. 바이오시밀러의 언급은 기간 2부터 나타났지만, 생산에 대한 구체적인 내용을 기간 3에서 볼 수 있었다. 기간 4에서는 맞춤형 신약 개발의 비중이 비약적으로 증가하였다. 라이브러리, 데이터, 플랫폼 기반의 맞춤형/정밀의료 트랜드가 반영된 결과로 예상된다.

기간별 차이를 분석하면 단일 주제에 대한 세분화가 진행되었다. 기간 2의 백신(바이러스, 인플루엔자, 파킨슨병)의 경우 기간 3에서 항체(인플루엔자, 바이러스, 인간항체)와 백신(감염모델, 예방백신, 신약, 자궁경부암)으로 상세 분류가 되었다. 또한 세포치료제와 항체 치료제의 세분화가 급격히 이루어졌는데, 세포치료제의 경우 기간 3의 2개에서 기간4는 4개로 증가하였다. 기간 3의 '세포치료제 - 면역세포치료제' 관련 주제의 경우, T세포, CD200 등 상세한 표적 세포와 단백질이 확인되었으며, 기간 4에서는 면역세포 치료제, ADC 기술, CAR 등 연구 표적 단백질, 상세 기술의 표현 등이 발현되었다. 이는 동일 주제 내에서 시간의 흐름에 따라 연구 타깃이 변화되었다고 유추 할 수 있다. 주제에 대한 분화와 비중의 변화를 직관적 분석을 위해 색으로 표현하였다. 파란색은 항체, 백신 관련 연구 주제이고, 노란색은 세포치료제 관련 주제이다. 이외에 주제별 주요 연구 내용은 다음과 같이 함께 생성된 단어들로 관찰 가능하다. 한국인 관련 주제(간세포암, 관절염, 파킨슨), 합성신약(전임상, 다국적연구), 암 줄기세포(Cetuximab 억제제, BVAC 항암백신) 이는 현재 주제에서 어떠한 의학 분야에 연구가 집중되고 있는지 보여준다.

사회적 이슈 또한 본 자료에 표현이 되었는데, 백신관련 연구 주제를 분석하면, 기간 1(조류인플루엔자), 기간 2, 3(인플루엔자), 기간 4(메르스)와 같이 해당 기간에서 사회적으로 관심이 높았던 주제들도 확인 할 수 있었다. 분석 기간의 전체적인 추이는 대표적으로 항암제에 대한 연구가 13% 이상으로 전 기간에 높았으며, 기간 1, 2에서는 신약후보 물질 (천연물 등)을 탐색이 주요하였다면, 기간 3, 4에서는

표적 치료, 세포 치료제, 줄기세포로 이어지는 차세대 바이오 의약품의 연구로 옮겨져 가는 경향을 볼 수 있다. 합성 신약의 비중보다는 천연물, 바이오 의약품의 비중이 압도적으로 높으며, 바이오시밀러, CMO(위탁생산)은 기간 2, 3, 4에 걸쳐 유사 비중으로 나타난다. 인력, 생산, 등록, 표준화, 지적재산 등 신약 개발을 위한 인프라 관련 내용도 전 기간에 분포하고 있다.

LDA 분석은 연구자가 지정하는 주제의 수에 맞는 주제별 단어를 표시해주지만, 단어와 단어 간 표현의 상관관계가 반영이 되지 않는다. 이에 LDA2Vec을 활용하여, 주제 내의 단어 간 영향을 다음 실험 결과에서 보고자 한다.

제 3 절 LDA2vec 분석 결과

LDA2vec는 비지도 학습 알고리즘으로 지도 학습과는 상이하게 학습 결과가 정답과의 차이로 재 학습 되지 않는다. 이에 LDA2vec의 군집화 성능 확인이 필요 하며, 새로운 방식의 LDA2vec를 활용한 분석 방법과 결과에 대해 작성하고자 한다.

1. 군집화 성능

LDA2vec의 결과물은 LDA와 동일하게 주제와 그 구성 단어로 표현된다. LDA와 동일하게 단어들로 표현되는 결과는 연구자에 따라 분석이 상이할 수 있음으로, 군집화 특성과 LDA2vec의 손실 함수 성능 기준으로 하여, 최적의 분석을 하고자 하였다. LDA2vec의 손실 함수는 (1)과 같이 정의한다.

$$\mathcal{L} = \mathcal{L}^{d} + \sum_{ij} \mathcal{L}_{ij}^{neg} \qquad (1)$$

손실 함수는 Document Weight의 디리클레 우도함수와 Skipgram Negative Sampling Loss의 합으로 표현되며, 우도 함수는 학습 진행 시, 희소성을 유지하여 문서 간의 차이를 극대화하고자 하고, 우변 또한 단어와 주제 벡터를 동시에 학습하게 된다. 이를 Tensorboard(그림 4-5)를 통하여 변화 추이를 확인 할 수 있으며, 손실 함수가 낮아 짐에 따라 주제와 주제 간 그리고, 주제 안에서의 단어의 중복성이 개선됨을 볼 수 있었다.

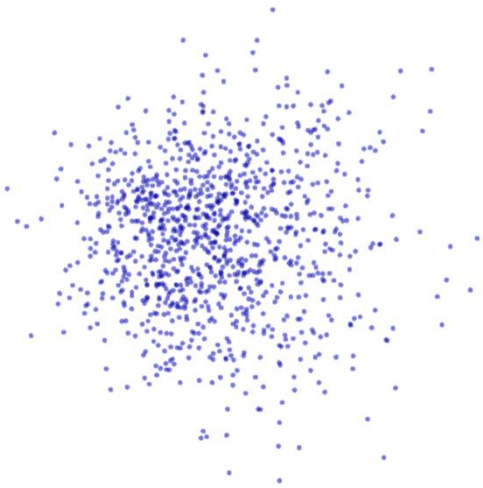
하지만, Epoch이 늘어나며 학습이 지속되고, 손실함수가 낮아진다고 하여, 연구자가 분석하기에 용이하게 변화 주제와 단어의 구성이 이루어지는 것은 아니다. 손실 함수가 0.6 이하로 수렴되는 경우도 있었지만, 손실 함수가 1.5 이하로 수렴되고, epoch의 수가 500 epoch이 넘어가면 유사한 성능을 확인 할 수 있었다. 이는 아직 군집화를 기반으로 한 비지도 학습의 특성이라 볼 수 있겠다.

[그림 4-5] LDA2vec Loss의 Epoch에 따른 변화

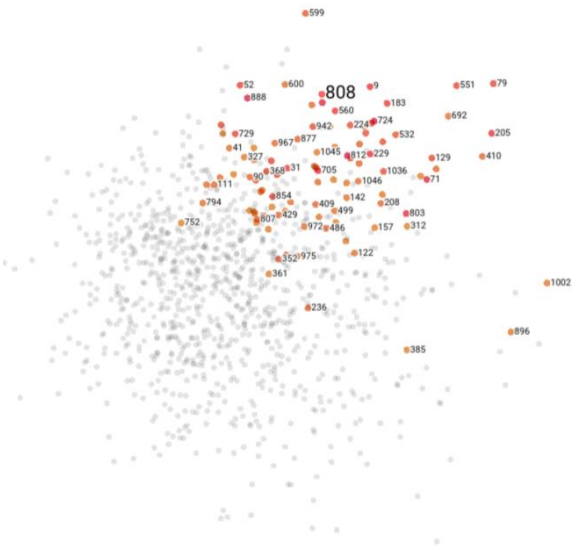
(Google Tensorboard 활용)

손실 함수 확인 이후, 본 연구의 목적인 효율적인 군집화를 통한 분석을 시각적으로 확인하기 위해 대표적 차원 축소 방법인 PCA (Principal Component Analysis)와 t-SNE (t-Stochastic Neighbor Embedding)을 활용하여, 결과 데이터의 군집화를 확인하였다. 그림 4-6은 LDA2vec의 단어 임베딩 결과를 PCA로 2차원에 표현한 그림이다. 단어 임베딩이 편중되지 않고 학습됨을 확인 할 수 있다. 나아가 주제에 단어들의 군집화를 확인하고자 하여, 동일 주제의 단어만 그림 4-6에서 붉은색으로 표현하였다. 이에 대한 결과는 그림 4-7에서 확인 할 수 있다. 그림 4-7에서 같은 주제의 단어들이 우상향 위치에 포진하며 군집화되는 모습을 볼 수 있다. 다른 주제들 또한 비슷한 유형의 군집화를 이루며, 다른 위치에서 표현됨을 확인하였다. 이는 군집화 학습이 되었음을 알 수 있다. 다른 방법으로도 군집화를 확인하였는데, 그림 4-8은 t-SNE를 적용하여, 단어의 군집화를 실 단어와 함께 표현한 그림이다. 그림 4-8에서도 주제 주변으로 주제 벡터와 유사한

단어들이 포진됨을 알 수 있다. LDA2vec의 주제의 수는 LDA와 동일하게 연구자가 지정하는 하이퍼파라메터이다. LDA2vec 분석 초기

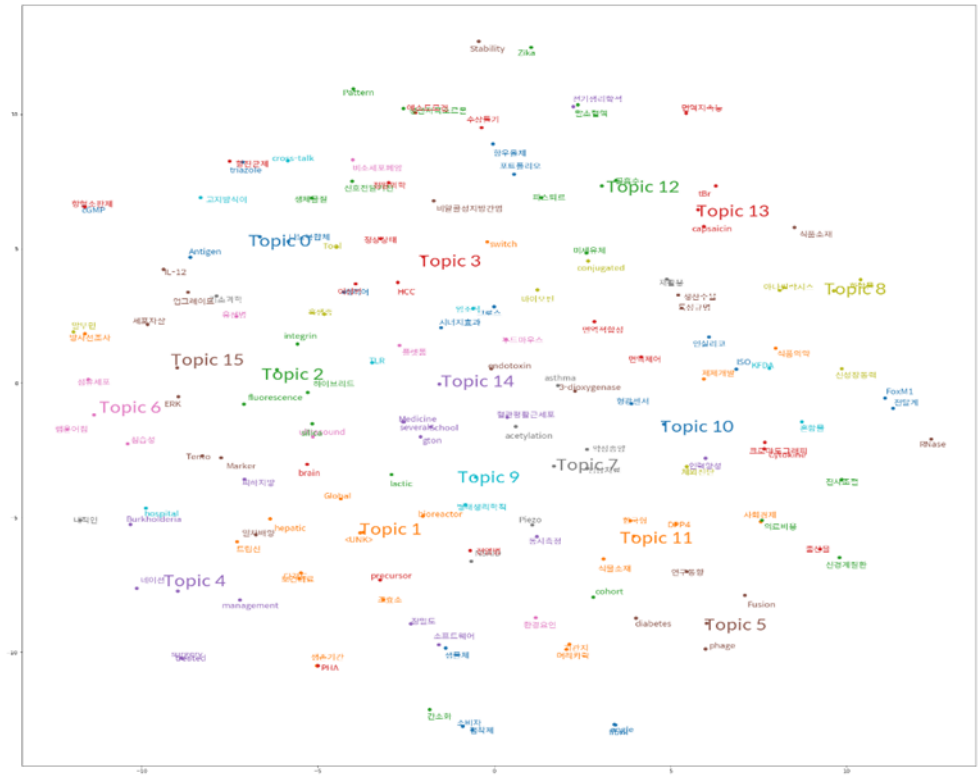


[그림 4-6] 단어 임베딩 PCA 차원 축소, 기간 4



[그림 4-7] 단어 임베딩 PCA 군집화 확인, 기간 4

주제 수는 기존 LDA의 주제 수와 동일 하게 진행을 하여 성능을 비교하였다. 표4-8은 2015~2017년까지 16개의 주제를 기준으로 실행한 LDA2vec이다. (실행 조건: epoch: 500, lambda: 200, alpha: 0.04, learning rate: 0.005, switch loss epoch: 20) 기존 LDA와 상이한



[그림 4-8] t-SNE 활용 주제 별 단어 표현, 기간 4

점은 주제에 대한 해석이 LDA의 경우 단어의 조합으로 유추해야 했다면, LDA2vec의 경우, 주제를 설명 할 수 있는 대표 단어가 분석 주제에 포함된 경향이 확인된다. 표 4-8의 첫 번째 주제명으로 선정한 '유전자 치료제'의 경우 유도 기술, 단백질, 싸이토카인(면역세포분비단백질), NF-kB(세포 생존 조절 인자)등의 단어들이 유전자 치료제와 연관이 높아 신약의 한 종류가 대표적으로 표현이 되었고, 다섯 번째 주제인 발암 기전 역시 역분화, 신호전달기전, Sox2 등을 포함한 대표적 단어로 분석 가능하였다. LDA2vec의 표현은 기존의 단어 빈도수 기반의 모델링이 아닌, 단어 임베딩과 주제 임베딩의 학습을 통해 유사 임베딩 벡터를 가진 단어들이 코사인 유사도를 기반으로 표현되어 대표 단어를 선정하는데 용이 하다 판단된다. 더욱이 표현 단어들도, 빈도수에 절대적인 영향을 낮게 받아 주제와 관련이 높은 전문 단어의 표현 또한 관찰된다. 예로 표 4-8의 주제3에 포함된 '자기공명영상'의 경우 전처리 과정에서 진행한

토픽	주제	단어	단어	단어	단어	단어
1	유전자치료	유전자치료,	유도기술,	프로스트,	제브라피쉬,	NF-kB,
		단시간,	vehicle,	proteins,	싸이토카인,	식물소재
2	소재화	임성학,	금나노입자,	식물추출물,	대사물,	Pharmaceutical,
		골분화,	Biofilm,	소재화,	communication,	치주질환
3	N/A	MRI,	유지시,	자기공명영상,	aptamer,	제품생산,
		소화효소,	SHP-1,	신장질환,	LPA,	que
4	신경성장인자	유일하,	결정학,	순수분리,	아토피,	분석장비,
		fluorescence,	신경성장인자,	resonance,	channel,	biochemical
5	발암기전	역분화,	호환성,	<UNK>,	Sox2,	인실리코,
		신호전달기전,	Oral,	center,	발암기전,	한국형
6	신경질환	유용물질,	열발생,	PSA,	DMF,	환자유래동물모델,
		창의성,	뽕나무,	신경질환,	바이오틴,	알고리즘
7	줄기세포	G-CSF,	PMS,	오프라인,	항산화제,	산학연,
		업그레이드,	세포구,	독성작용,	외적인,	Stem
8	암발생	repair,	분화치료,	클래스,	크로마토그래피,	암발생,
		아시네토박터,	화학유전체,	생체모방화학,	GSK,	리서치
9	발암기전	alternative,	transduction,	Hub,	ERK,	생체막,
		발암기전,	Artemether,	완제품,	aroma,	Elim
10	유전질환	타액선,	재활용,	상동성,	heterogeous,	유전질환,
		나노복합체,	혈관치료,	암억제,	가톨릭대,	대사관
11	T세포치료제	SAM,	pyruvate,	소프트웨어,	아미노기,	국가경쟁력,
		eva,	ult,	in-vitro,	colon,	T세포치료제
12	정밀치료	NGS,	Translation,	림프절,	치료요법,	생체적합성,
		복제돼지,	대동물,	인력양성,	heterogeneous,	nasal
13	N/A	비대칭합성,	선유도,	디지털,	Ca2+,	음낭수종절제술,
		EMT,	내시경,	대체제,	한국형,	중금속
14	뇌종양줄기세포	국제규격,	Heterologous,	MMP,	phenotyping,	mitochondria,
		호르몬,	sodium,	국제공동연구,	NGS,	뇌종양줄기세포
15	N/A	업그레이드,	상호결합,	walls,	의료기기,	의료기,
		기초과학,	시상하부,	Gene,	항균단백질,	stream
16	전사조절인자	<UNK>,	전사조절인자,	피로감,	결합능,	CoV,
		trypsin,	복합제,	가용성,	surgery,	treated

[표 4-8] LDA2vec 주제 16개 적용, 기간 4

NPMI의 결과로 조합된 문서 특화 단어이며, LDA 분석에서는 표현되지 않았던 단어이다.

기간 4에 대해, 16개의 주제로 모델링을 진행했을 경우, 일부 주제 내의 단어가 상호 연관성이 높다고 판단되지 않아, 3, 13, 15번의 주제명을 산정하기 어려웠다. 이를 극복하기 위해 주제 수를 높여 주제를 상세화하여 해석을 용이하게 하고자 하였다. 주제 수는 동일 기간에 대해여, LDA 진행 시 계산되었던 Coherence Score를 참고하여 22개로 확장하여 실험을 진행하였고, 불용어의 경우 최종 불용어 처리의 수준보다 낮게 하여, 대표 단어의 민감도를 높여 성능의 차이를 보고자 하였다. 그 결과 주제 4번에 대한 주제 선정은 대표 단어의 부재로 선정이 어려웠지만, 그 외 21개의 주제에 대해서는 분석 결과가 제공한 단어 중 대표 단어의 표현이 강화되어 주제 선정이 가능하였다.

토픽	주제	단어	단어	단어	단어	단어
1	천연물소재	therapeutic,	생체모사,	Cytokine,	중추신경계,	피드백,
		cannabinoid,	ISO,	Index,	천연물소재,	경조직
2	형질전환체	Vol,	영상진단,	보건학,	sue,	tive,
		형질전환체,	메틸화,	Double,	천연물,	임상결과
3	면역제어	DMPK,	재활용,	생체표지자,	유병률,	면역제어,
		카세트,	UCP,	공중보건학,	synergism,	표준화
4	N/A	capsaicin,	당뇨병,	피드백,	항결핵,	각질형성세포,
		chemokine,	빗살수염벌레,	Porous,	방사선조사,	다차원
5	화학치료법	가수분해,	업그레이드,	매개인자,	생체모방화학,	화학치료법,
		생체신호,	LIGHT,	분해과정,	Library,	세포소기관
6	독성연구	MicroRNA,	나노다공체,	치주질환,	예후인자,	interface,
		IL-12,	독성연구,	fexar,	회충증,	난발현성
7	면역성	형태형성,	SLAMF,	단핵구,	면역성,	rectly,
		국가경쟁력,	Diazadiphosphocine	나노복합체,	섬유세포,	분자간
8	lipoprotein, (콜레스테롤)	치조골,	광영상,	Signal,	공중보건학,	무산소,
		영상유도,	보툴리눔,	lipoprotein,	Kru,	콜레스테롤
9	antagonists, (길항제)	development,	Collagen,	음경해면체,	이차대사산물,	나노구조체,
		일상생활,	품질향상,	antagonists,	비세포,	DNA
10	Regulator	<UNK>,	나노스케일,	Regulator,	transgenic,	secretion,
		고지질혈증,	BBB,	CYP4A,	에세이,	PTH
11	분자유전학	분자구조,	유전독성,	exosome,	분자유전학,	뇌신경질환,
		HSF1,	proline,	angiography,	인공지능,	동식물
12	장기이식	장기이식,	ALT,	간암모델,	bol,	비수술,
		진단키트,	JAK,	리피도믹스,	automatic,	경제성
13	메타볼로믹스	ISO/TC276,	임상효과,	방사선민감도,	NAD,	insulin,
		염증동물모델,	colitis,	SEB,	메타볼로믹스,	산학협력
14	퀩티도미메틱 페노믹,	mitochondria,	퀩티도미메틱,	healing,	자폐증,	수치화,
		페노믹,	Chro,	정밀도,	ium,	창의성
15	생체신호	시험평가,	고지방식이,	Image,	휴지기,	분석장비,
		시신경,	뇌세포,	건강기능식품,	오프라인,	생체신호
16	Bioinformatics	정맥투여,	proteomics,	환기시설,	가시화,	농축기법,
		exi,	phase,	일체형,	소프트웨어,	Bioinformatics
17	혈청학	단시간,	임상치료,	RANK,	천포창,	혈청학,
		radiation,	엔도테린,	생체시료,	시너지효과,	Index
18	화장품	Hydantoin,	디바이스,	마이크로니들,	배양세포,	WNT,
		노령화,	골격근,	Pharmacophore,	화장품,	exi
19	항암면역세포	절돌연변이,	생물체,	항암기전,	Antisense,	엠스타인바바이러스,
		Reph,	colon,	Flavonoid,	고민감도,	항암면역세포치료
20	인공지능	인공지능,	유전병,	에리스로포이에틴,	무척추동물,	Oral,
		Probiotics,	transduction,	포화지방산,	치료표적,	MEK
21	이온통로	세뇨관,	병용처리,	TLC,	interface,	산학협력,
		ERK,	사회경제,	이온통로,	in-vitro,	Bacillus
22	diabetes	diabetes,	Hedgehog,	조추출물,	normal,	마이크로입자,
		혼합물,	transgenic,	rvum,	Proliferation,	바이오산업

[표 4-9] LDA2vec 주제 22개 적용, 기간 4

LDA는 16개의 주제를 해당 데이터에 적합한 주제의 수로 판단하였지만, LDA2vec에서는 대표 단어의 표현력 강화로 주제의 수를 22개로 확장하여 세분화된 분석이 가능하였다.

2. 상세 주제 탐지 및 예측 결과

2.1. 타깃 단어의 주제 안에서의 변화

LDA2vec의 군집화 성능 확인 이후, LDA 분석과 동일한 전처리 진행 과정을 진행하였다. LDA는 기간별 주제 분석 이후 동일 주제에 대해 연결함과 동시에 문서내의 비중의 변화를 탐지하고자 하였다면,

LDA2vec의 분석은 변화를 보고자 하는 타깃 단어가 주제 안에서 다른 단어들의 영향을 받으며 변화하고, 또한 시간의 흐름에 따라 영향을 받는 단어들의 차이를 통해, 주제의 변화를 분석하고자 하였다. 본 분석은 동일한 단어들의 표현이 다른 기간의 주제에서도 일치하는 비율이 높다면 같은 의미를 가지는 주제로 볼 수 있다는 가설로 시작하였다. 군집화에서 나타난 주제의 세분화 성능을 통하여, LDA보다 증가한 주제의 수로 LDA2vec을 진행하였다. 기간별 주제 수는 기간 1(17개), 기간 2 (19개), 기간 3(19개), 기간 4(22개)로 지정 하였다. 모두 LDA 진행 시 Coherence Score의 정상치 부근에 포함된다. 모든 기간에 대한 실행 조건은 epoch: 500, lambda: 200 alpha: 0.04, learning rate: 0.005, switch loss epoch: 20으로 진행 하였다. 실행 결과는 Appendix에 추가하였다.

타깃 단어의 기간별 변화를 확인하기 위하여, 임의의 타깃 단어를 선택하였다. 첫 번째로 변화를 확인하고자 하는 타깃 단어는 '사이토카인'이다. '사이토카인'은 면역 세포에서 분비되어, 다른 세포에 영향을 주는 모든 단백질을 통칭한다. 본 단어는 LDA2vec을 진행한 기간 1에서 3, 11, 13, 16번 주제에 포함되어 있다. 총 4개의 주제에서 타깃 단어의 거리를 유클리디안 거리로 계산하였다. 유클리디안 거리는 Tensorflow(Pairwise Euclidean Distance)을 활용하여 두개의 tensor의 거리를 계산 하였다. 주제를 표현 하는 데 있어 주제와 가까운 단어가 주제에 높은 영향을 미치는 방향으로 학습을 하였기에, 타깃 단어와 주제의 거리가 가장 짧은 주제를 기준으로 해석을 시작하였다. 기간 1에서 '사이토카인'의 영향을 가장 많이 받은 주제는 13번 주제로, 거리는 1.255로 나타난다. 기간 2에 대해서 또한 LDA2vec을 진행하였고, 기간 2에서 '사이토카인'을 보유한 주제는 18번 하나로 확인되었다. 기간 3에서의 '사이토카인'을 보유한 주제는 총 4개로 5번, 10번, 11번, 16번으로 나타났으며, 이번에는 주제와 '사이토카인'의 거리를 사용하지 않고, 기간 2의 18번 주제와 기간 3의 5번, 10번, 11번, 16번 주제의 단어 일치도를 계산하였다. 이는 본 결과분석 서두에 언급한 '주제를 구성하는 단어의 일치율이 높다면, 같은 의미를 가지는 주제이다'는 가정을 적용하였다. 결과 기간 2의 주제 18번은 기간 3의 5번 주제와 일치율이 16.3%로 가장 높았고, 동일한 방법으로 기간 4의 주제들을 기간 3의 5번 주제와 비교하여 기간 4의 3번, 19번 주제와 가장 높은 일치도(14%)를 얻을 수 있었다. LDA도 주제별 단어를

기간1 – 주제 13			기간2 – 주제 18		
순서	단어	주제와 거리	순서	단어	주제와 거리
0	면역내성	1.254133	0	폴리에틸렌글리콜	1.250906
1	anticancer	1.254141	1	Acinetobacter	1.250922
2	사회인지기능	1.254153	2	Cornell	1.250978
3	ginsenoside	1.254556	3	methylation	1.251157
4	spheroid	1.254591	4	알고리즘	1.251165
5	의약합성	1.254606	5	패치형	1.251234
6	engineering	1.254656	6	재정립	1.251292
7	FK506	1.254762	7	포유류	1.251337
8	혼합백신	1.254899	8	블록공중합체	1.251399
9	혈관형성	1.254951	9	chelat	1.251416
10	사이토카인	1.255004	10	사이토카인	1.251469
11	위장장애	1.255012	11	Resonance	1.251515
12	patent	1.255141	12	nAChR	1.25154
13	MRI	1.255168	13	측정기	1.251572
14	면역치료백신	1.255184	14	Roche	1.251587
15	화학유전체학	1.255204	15	novel	1.251636
16	외용제	1.255209	16	Pichia	1.251722
17	면역능	1.255264	17	항우울증	1.251867
18	지방산	1.255569	18	elim	1.251918
19	maleate	1.25557	19	이식면역	1.251984
20	항산화단백질	1.255732	20	양성전립선비대증	1.251999

기간3 – 주제 5			기간4 – 주제 3			기간4 – 주제 19		
순서	단어	주제와 거리	순서	단어	주제와 거리	순서	단어	주제와 거리
0	eudomallei	1.295955	0	GTP	1.333566	0	배양법	1.360518
1	전산화	1.296207	1	음전하	1.333606	1	cyclosporin	1.360519
2	산학연	1.296214	2	황색포도알균	1.333629	2	석회화	1.36052
3	인수공통전염병	1.296227	3	거핵구	1.33368	3	황화수소	1.360525
4	세계보건기구	1.296389	4	Spike	1.333754	4	비부비동염	1.360558
5	Hardware	1.296402	5	Embryo	1.333754	5	번역개시인자	1.360569
6	혈액응고	1.29642	6	나노소재	1.333755	6	전자현미경	1.360572
7	ODN	1.296487	7	막전압	1.333907	7	스페로이드	1.360577
8	언어기	1.296603	8	metabolism	1.333929	8	건강기능성	1.360594
9	마취제	1.29666	9	알츠하이머	1.333955	9	Smad2	1.360628
10	사이토카인	1.29674	10	사이토카인	1.334224	10	사이토카인	1.360631
11	상처치유	1.296768	11	zedox	1.334244	11	구충제	1.360703
12	암세포	1.29678	12	phosphatid	1.334363	12	STING	1.360709
13	신경퇴행성질환	1.296802	13	항염효과	1.334378	13	가상인체모델	1.360719
14	당전이	1.296812	14	무척추동물	1.33438	14	사이클로스포린	1.360765
15	약제내성	1.296851	15	영상진단	1.334391	15	rst	1.360789
16	류비저	1.296853	16	cardiac	1.334617	16	아세테이트	1.36084
17	인간항체	1.296903	17	비타민D	1.334624	17	Notch1	1.360849
18	상처치료제	1.296905	18	MRAB	1.334742	18	척수염	1.360884
19	유기물	1.29697	19	항혈전	1.334754	19	카세트	1.36094
20	CMO	1.296993	20	scale	1.334756	20	호산구	1.360941

[표 4-10] 타깃 단어(사이토카인)의 기간 별 주변 단어의 변화

활용하여, 유사한 방법으로 기간별 유사 주제들을 연결 할 수 있다.

표 4-10은 상기에서 진행된 방법으로 연결된 기간별 동일 주제와 변화를 보고자 한 타깃 단어를 중심으로 표현한 표이다. 타깃 단어 사이토 카인 주변 10개 단어의 변화를 기간에 따라 보여준다. 기간 전체를 통해 면역 관련 내용들이 사이토카인과 연관성을 가지며,

질병의 경우, 암세포, 알츠하이머등과 관계를 보인다. 기간 4에서 SMAD2(Mothers against decapentaplegic homolog 2, TGF-β의 신호를 중재하며, 세포의 증식, 자멸, 분화 등의 활동을 다양하게 관장)에 최 근접성을 나타내는데, 이는 최근 기간인 4에서 사이토카인은 SMAD2와 연관성 높게 연구가 진행되었을 가능성이 높음을 유추 할 수 있다. 본 연구에서는 이처럼 기간별 동일 주제를 탐색하고, 주제 안에서 특정 단어 주변의 변화를 관찰 하여 영향력이 높은 단어들을 확인하였다. 동일 주제 내에 동시에 존재하는 단어의 변화 또한 주제를 상세 분석하기 위한 요소이다. 이에 다음 항에서는 주제를 중심으로 두 개의 상이한 기간에서 동일한 단어의 움직임을 보고자 한다.

2.2. 타깃 단어의 주제 안에서의 기간별 변화

2.1항은 주제 내의 타깃 단어 주변 단어의 변화를 관찰하였다면, 2.2항에서는 기간이 다른 동일 주제에서의 단어의 변화를 보고자 하였다. 이를 위해 단어와 주제와의 거리순으로 순위를 부여하여, 두 기간에 대해여 동일한 단어의 변화를 탐지하였다. 대상 기간은 3(2012~14), 4(2015~2017)로 선정하고, 주제와 타깃 단어는 2.1항과 동일하게 선정하였다.

순위	차이	기간3	기간4	단어	순위	차이	기간3	기간4	단어
0	-130	139	9	신경전달물질	20	-71	130	59	glycan
1	-124	134	10	ilure	21	-67	73	6	기초과학
2	-124	138	14	서구화	22	-61	93	32	YAP
3	-109	116	7	TGFBIp	23	-59	114	55	대장염
4	-108	126	18	병용효과	24	-58	94	36	막단백질
5	-100	123	23	신경병증	25	-54	54	0	샤페론
6	-99	127	28	비마약성	26	-52	77	25	당사슬
7	-94	132	38	당뇨치료제	27	-50	121	71	숙주세포
8	-93	98	5	mono	28	-49	62	13	성체줄기세포
9	-91	131	40	metabolism	29	-45	89	44	반세포
10	-90	135	45	에이즈	30	-42	83	41	사이토카인
11	-89	90	1	연수기	31	-42	50	8	항박테리아
12	-86	125	39	GTP	32	-40	88	48	optive
13	-85	120	35	항균성	33	-38	137	99	ultrasound
14	-79	106	27	마우스모델	34	-37	59	22	nucleotide
15	-75	105	30	vascular	35	-36	111	75	건강보험
16	-75	136	61	심혈관질환	36	-34	128	94	활성연구
17	-74	86	12	전임상실험	37	-31	35	4	β-catenin
18	-74	100	26	의과대학	38	-28	43	15	epitope
19	-74	108	34	HDAC	39	-27	101	74	선도화합물

[표 4-11] 기간 3에서 기간 4의 동일 주제 내의 주제 벡터로부터 단어 벡터의 순위 변동

표 4-11는 '기간 3'과 '기간 4'에서의 동일한 단어가 각각의

주제로부터 거리의 차이를 표현하였다. 타깃 단어인 사이토카인은 기간 3(순위89) 대비 기간 4(순위44)에서 주제와 가까워짐을 보였고, 사이토카인 단일 단어로 2.1항에서 분석한 결괏값과 일치하는 단어는 GTP와 Metabolism로 확인되었다. GTP와 Metabolism 또한 각각 86, 91순위가 상승하여 주제와 근접하게 되었다. 이는 타깃 단어(사이토카인)의 기간별 이동을 관찰함과 동시에, 타깃 단어(사이토카인)과 관련된 단어(GTP, Metabolism)의 움직임 또한 관찰 가능하여, 기존 LDA에서 확인 불가능 한 단어와 단어 간의 관계를 확인하였다. 주제 전체적으로는 신경전달물질, TGFBIp, 신경병증 등의 단어들이 주제의 중심으로 이동함을 볼 수 있었다. 이는 다른 기간의 동일 주제로 선정하였지만, 주제 내부에서의 비중의 변화를 탐지 할 수 있어, 주제 해석 시, 우선순위의 조정을 해야 할 정보를 제공한다.

최근 관심이 급격히 높아진 인공지능 관련 연구의 경우, 기간 1, 2, 3에서 인공지능 단어를 찾아볼 수 없었다. 하지만, 표 8-1를 확인하면 기간 4, 주제 4번에서 3위의 표현력으로 주제에 포함되어 있다. 주제 내에서 동시 표현 단어로는 바이오산업, 올리고뉴클리오타이드, 디지털이며, 인공지능과 주변에 동시에 발현되는 단어는 표 4-12와 같다. 이는 인공지능 활용 가능성이 있는 바이오 연구 분야로 추측 할 수 있다.

순서	word	topic
0	임상결과	1.310215
1	PPAR	1.310486
2	바이오센서	1.310515
3	conformation	1.310535
4	미래전략	1.310789
5	진단마커	1.310867
6	펩티도미메틱	1.310898
7	LDL	1.311068
8	NRPS	1.311294
9	표면항원	1.311421
10	인공지능	1.311589
11	cross-talk	1.311772
12	나노바이오센서	1.311947
13	단백질치료제	1.311958
14	data	1.311973
15	CXCR4	1.312012
16	brain	1.312016
17	유전정보	1.312098
18	Argonaute	1.312171
19	rever	1.312217
20	아시네토박터	1.312496

[표 4-12] 기간 4의 단어 인공지능과 유사 임베딩 벡터 단어

2.3. 트렌드 예측

간단한 선형 모델로 예측 모델을 구성해 보았다. 다양한 외부의 요인으로 인해 예측이 매우 어려운 증시, 국제정세 등과는 다르게 상대적으로 제한적인 환경에서 연구되는 학문의 경우 단기간 예측의 가능성을 가정하여 진행하였다. 예측은 기간별 동일 주제 안에서 타깃 단어 주변 단어의 변화를 관찰하는 것으로 한다. 기간별 동일 주제는 2.1항의 가정과 동일하게 단어의 일치율을 기준으로 주제를 연결하였다.

독립적인 예측을 위해, 학습 데이터를 3개 세트로 분리하였다. 기존 연구는 기간의 구분 없이 15만 건을 학습했던 것과는 상이하게, 연구 제안서의 BT 데이터를 기존과 동일한 기간 1 (30,715건), 기간 2 (32,444건), 기간 3(37,969건)으로 분리하여 각각 학습 진행하였다. 이후 학습된 데이터는 다음의 기간에서 LDA2vec의 사전 학습 데이터로 활용하여, 학습 데이터와 분석 데이터의 연관성을 최소화하고자 하였다. 이에 표 4-13과 같이 연속적으로 학습과 분석을 진행하였다.

기간	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
기간 2	Training set			Reference								
기간 3				Training set			Reference					
기간 4							Training set			Test set		

[표 4-13] 기간 별 데이터 활용 구조

우선 표 4-13의 Reference로 표기된 기간 2, 3의 LDA2vec 결과로부터 분석하고자 하는 타깃 단어와 주변 단어의 유클리디안 거리를 각각 계산하였다. 이는 표 4-14에 '기간 2', '기간 3'으로 표기되었다. 이후, 기간 2, 3의 동시 표현된 단어의 움직임을 타깃 단어를 중심으로 확인하고자 하였다. 기간 2에서 3으로 시간이 흐름에 있어 타깃 단어 주변 단어가 타깃 단어로 이동을 기간 2와 3에 걸쳐서 한다면, 다음 기간 4에서도 더 근접할 것이라 예상하였다. 이후 기간 4의 LDA2vec의 결과와 비교하여, 예측의 정확도를 분석하였다.

분석 타깃 단어는 '바이오시밀러'로 선정하여 1차 진행을 하였다. 그 결과 표 4-14와 같은 결과를 얻을 수 있었다. (기간 2의 12번째, 기간 3의 0번째, 기간 4의 5번째 주제가 가장 높은 단어 일치도를 보여 동일 주제로 가정하였다.) 표 4-14의 '기간 3, 2 차이'는 '기간 3'에서 '기간 2'를 단순 차이를 구한 값으로, 주변 단어가 타깃 단어를 향한 움직임을 나타낸다. '-'의 차이를 보이는 단어들은 타깃 단어와

가까워지는 단어이고, '+'의 차이를 보이는 단어들은 타깃 단어와 멀어지는 단어들이다. 이후 테스트 '기간 4'와 동일한 분석을 '기간 3'과 진행했을 때, '바이오시밀러'의 경우 총 25개 동일 단어에서 15개의 (60%) 의 예측 일치율을 보였다. 이는 기간 2와 3을 분석 했을 때와 같은 방향성을 기간 4에서도 나타냄을 의미한다. 표 4-14를 분석하면, '바이오시밀러'의 경우 Manufacturing (생산), 프로젝트, 식품의약과 같은 생산/관리/규격과 같은 내용으로 집중된다 할 수 있고, 이외의 단어들은 관련도가 멀어진다 할 수 있다.

단어		기간 2	기간 3	기간 4	기간 3, 2 차이	기간 4, 3 차이	일치 여부
1	Manufacturing	1.3323	1.3206	1.1270	-0.0118	-0.1936	o
2	프로젝트	1.3695	1.3593	1.3398	-0.0102	-0.0196	o
3	식품의약	1.4012	1.3346	1.3297	-0.0666	-0.0050	o
4	보건의료	1.2534	1.3668	1.3670	0.1134	0.0002	o
5	질병모델	1.2699	1.2862	1.3102	0.0163	0.0241	o
6	환자유래	1.3225	1.3767	1.4091	0.0542	0.0325	o
7	다기능성	1.3429	1.3884	1.4218	0.0456	0.0334	o
8	뇌손상	1.1578	1.3496	1.4067	0.1917	0.0571	o
9	VEGFR	1.0874	1.2338	1.3059	0.1464	0.0722	o
10	식생활	1.2881	1.3193	1.4064	0.0313	0.0870	o
11	식약청	1.2848	1.3307	1.4245	0.0459	0.0938	o
12	당뇨치료제	1.1449	1.3469	1.4459	0.2020	0.0989	o
13	유전학	1.3443	1.3595	1.4930	0.0152	0.1335	o
14	자궁내막	1.2616	1.3505	1.4962	0.0889	0.1457	o
15	시뮬레이션	1.2116	1.2680	1.5344	0.0564	0.2664	o
16	약효평가	1.3380	1.3413	1.2180	0.0032	-0.1233	x
17	비교동등성	1.2318	1.3171	1.2174	0.0853	-0.0997	x
18	장기이식	1.3244	1.3830	1.3327	0.0586	-0.0504	x
19	항균효과	1.2895	1.3169	1.2687	0.0274	-0.0482	x
20	유전자형	1.3962	1.3964	1.3536	0.0001	-0.0428	x
21	줄기세포치료제	1.2527	1.3518	1.3220	0.0991	-0.0298	x
22	toxicity	1.2699	1.3583	1.3314	0.0885	-0.0269	x
23	림프관	1.3380	1.2880	1.3502	-0.0500	0.0622	x
24	공정개선	1.2729	1.1378	1.2191	-0.1351	0.0813	x
25	약물상호작용	1.3075	1.0961	1.3134	-0.2114	0.2173	x

[표 4-14] 타깃 단어 '바이오시밀러' 주변 단어 변화 예측 결과

2.3 '트렌드 예측'의 서두에서 언급한 것과 같이, 간단한 선형 모델로 다른 타깃 단어를 선택하여 동일한 분석을 진행했을 시, 50% 이상의 예측 결과를 얻지 못하였다. 표 4-15는 타깃 단어 HDAC (Histone Deacetylases)를 중심으로 분석한 결과이다. 24%의 낮은 적중률을 보인다. 이에 예측에 대한 신뢰도는 현재 기간과 거리의 차이로 분석한 단순 모델의 한계라 할 수 있다. 하지만, 주제 내에서 타깃 단어의 움직임을 수치로 확인 할 수 있으며, 기간 간의 차이로 얻어진 수치의 '+', '-'의 조합은 3가지로 분리하여 분석 할 수 있는

단어		기간 2	기간 3	기간 4	기간 3, 2 차이	기간 4, 3 차이	일치 여부
1	아쥬번트	1.2442	1.3966	1.4707	0.1525	0.0740	o
2	항당뇨	1.2487	1.4059	1.4372	0.1572	0.0313	o
3	인력양성	1.2574	1.2734	1.4296	0.0160	0.1563	o
4	항노화	1.3047	1.3102	1.3290	0.0054	0.0189	o
5	약효평가	1.3330	1.3373	1.3893	0.0042	0.0521	o
6	치료약제	1.3139	1.3332	1.4562	0.0193	0.1230	o
7	RNAi	1.3054	1.1881	1.4831	−0.1174	0.2950	x
8	동위원소	1.3319	1.4324	1.3005	0.1005	−0.1319	x
9	근육세포	1.3102	1.2968	1.4235	−0.0134	0.1267	x
10	저분자화합물	1.3418	1.3528	1.3166	0.0110	−0.0362	x
11	전임상후보물질	1.2763	1.3237	1.2499	0.0474	−0.0738	x
12	재조합단백질	1.2977	1.2886	1.3498	−0.0090	0.0612	x
13	질환동물	1.3892	1.3397	1.4040	−0.0495	0.0643	x
14	산학연	1.3225	1.3896	1.3782	0.0671	−0.0114	x
15	국제공동연구	1.4449	1.2896	1.3554	−0.1554	0.0658	x
16	키메라	1.2725	1.2610	1.3077	−0.0115	0.0467	x
17	바이오산업	1.3250	1.4631	1.2875	0.1382	−0.1757	x
18	올리고뉴클레오티드	1.2609	1.2915	1.2574	0.0306	−0.0340	x
19	phosphatid	1.3428	1.3012	1.3464	−0.0416	0.0452	x
20	골격근	1.2719	1.2224	1.2673	−0.0495	0.0449	x
21	자생식물	1.3836	1.2851	1.4297	−0.0985	0.1446	x
22	시판허가	1.3736	1.3281	1.3887	−0.0455	0.0606	x
23	KFDA	1.3328	1.3922	1.3650	0.0595	−0.0273	x
24	tRNA	1.3067	1.2709	1.4157	−0.0358	0.1448	x
25	마이크로니들	1.3179	1.2964	1.4194	−0.0215	0.1230	x

[표 4−15] 타깃 단어 'HDAC' 주변 단어 변화 예측 결과

가능성을 제공한다. 1) '−', '−'의 조합은 강하게 타깃 단어 주변으로 이동하는 단어로 높은 연관성을 표현하고, 2) '+ ', '+ '의 조합은 1)와 반대로 타깃 단어에서 멀어져 중요도가 낮아짐을 나타낸다. 3)은 '+', '−'이나 '−', '+' 의 구성으로, 방향성을 알기 힘들다. 또한, '+', '−'이나 '−', '+'의 해석의 의미가 순서에 따라 다를 수 있다.

적중률과 상기 3가지의 상관관계를 보았을 때, 바이오시밀러처럼 단어 이동 방향의 적중률이 높을수록 연구 및 관련 분야의 진행 방향성이 단어들의 상관관계로 분석 가능하다. 하지만, HDAC와 같이 적중률이 낮으면, 방향성에 대해 판단하기 어렵다. 본 부분은 추후 연구에서, 주제의 생성과 소멸과 함께 단어의 분포의 관계를 관찰 가능할 때, 보다 정확한 판단을 할 수 있을 것으로 예상한다. 짧지만 두 기간(기간 2, 3)의 데이터의 차이를 기준으로 다음 '기간 4' 를 예측해 보았다. 높은 성능을 나타내지는 못했지만, 시간이 지남에 따라 타깃 단어 주변 단어의 영향력 차이를 분석 할 수 있었고, 주제가 가지는 방향성의 의미를 확인할 수 있었다.

제 5 장 결 론

학문적 지식이 빠르게 세분화되고 고도화됨에 따라, 연구자들은 연구 분야를 효율적으로 분리하고 탐구하고자 하는 내용의 변화를 상세하게 분석하는 방법이 필요하게 되었다. 이를 위해 본 연구에서는 전문적 연구내용을 담고 있는 2006 년부터 2017 년까지의 '국가 연구 제안서' 중, BT 분야 신약개발기술 데이터를 기반으로 주제 탐색 및 주제 내에서의 단어 간의 관계 변화를 통해 상세 주제 분석을 시도하였다.

바이오 분야에 사용되는 전문용어에 대한 검출 능력을 높이기 위해 전처리 과정에서 NPMI(Normalized Pointwise Mutual Information)을 활용하여, 말뭉치 사전을 업데이트하였다. 이를 통해 Coherence Score 가 LDA(Latent Dirichlet Allocation) 분석 시, 동일 주제 수에서 높아짐을 확인하였다. 의미론적으로 주제 분리 성능이 향상됨을 볼 수 있었다. 이후 신약개발기술 동향 분석을 위해 두 가지의 다른 토픽 모델링 기법을 활용하였다. 첫 번째로 기존 LDA 를 활용하여, 4 개의 각기 다른 기간의 주제를 분석하고, 동일한 주제에 대해서 연결함과 동시에, 주제가 전체 문서에서 차지하는 비율을 시각화하여, 연구 분야의 변화를 상세하게 분석하고자 하였다. 항암제에 대한 연구는 국내 신약개발에서 가장 많은 비중을 차지하였고, 세포치료제와 항체치료제의 연구가 다양하게 분화됨을 관찰 할 수 있었다. 또한, 기간 4 에서 최근 바이오 산업계의 트렌드와 부합하는 맞춤형 신약개발, 데이터, 플랫폼 관련 주제의 생성도 확인 할 수 있었다. LDA 를 통해 상세 분석을 위한 전체적인 주제 숫자와 주제의 흐름을 분석하였다면, 전통적 토픽 모델링에 단어 임베딩 정보를 동시에 학습하는 LDA2vec 을 기본 분석 알고리즘으로 활용하여, 주제 내 단어들의 상관관계를 분석하고자 하였다. LDA2vec 의 성능 확인을 위해 군집화 결과를 시각적으로 표현하고자 하였으며, PCA 와 t-sne 의 차원 축소 방법을 통해 학습 결과를 표현 할 수 있었다. 주제명 선정 있어 LDA 는 단어들의 조합으로 주제명을 유추해야 했지만, LDA2vec 의 경우 한 주제에 표현된 단어는 유사한 임베딩 벡터를 가진 단어들로 주제를 대표할 수 있는 단어가 표현되는 경향을 보여 주제명을 표현하기에 용이하였다. 이후 LDA2vec 의 분석 결과를 활용, 3 가지의 방법으로 상세 분석을 시도하였다. LDA 와 상이하게 주제의 주제명을 임의로 선정하지

않고, 연구자가 관심 있는 단어를 지정하여, 관심 단어가 포함된 주제를 기준으로 기간별 분석 및 주변의 단어 변화를 확인하고자 하였다. 첫 번째 상세 분석 방법은 동일 주제 안에서 타깃 단어 주변 단어의 변화를 관찰하였다. 동일 주제는 단어의 일치도로 산출하였다. 일반 단어 임베딩 기법과 다르게 주제로 군집화된 집단 안에서 타깃 단어 주변 단어의 생성과 소멸을 통해, 해당 기간 주제 내에서 타깃 단어에 영향을 주는 단어들을 확인 할 수 있었다. 두 번째로 연속적인 기간에서 동일한 주제 내의 전체 단어의 움직임을 관찰하였다. 주제 벡터와 주제에 포함된 단어의 개별 벡터와의 거리를 근접한 순으로 순위를 부여한 이후 연속적인 다른 두 기간에서 동일 단어의 주제와의 거리에 대한 순위 변화를 분석하였다. 주제 벡터에 유사한 벡터를 보유한 단어가 주제를 표현하는 중요 단어로 가정한다면, 급격히 순위가 주제로 가까워지는 단어는 전 기간 대비 후 기간에서 더 중요한 표현력을 보유한다고 판단하였다. 이는 동일한 주제라고 하여도 설명이 다르게 될 수 있음을 나타낸다. 나아가, 첫 번째 분석 방법에서 표현된 타깃 단어와 주변의 유사 단어들 또한 두 번째 방법에서도 관찰이 가능하다. 이는 주제 안에서 단어의 중요도 변화를 관찰함과 동시에, 단어와 단어의 관계를 동시에 분석 가능하였다. 세 번째 1 차적 선형 모델로 주제 변화를 예측하고자 하였다. 학습 데이터와 분석 데이터를 분리하여 순차적으로 토픽 모델링을 진행하였다. 기간 2, 3 으로 기간 4 의 트렌드를 예측해보고자 하였다. 일부 단어에 대해서는 50% 이상의 예측력을 보였지만, 일부 단어는 예측 할 수 없는 20%의 성능을 보였다. 이에 단순 예측의 한계를 보였지만, 예측 분석 이후, 타깃 단어 주변으로 이동하는 단어와 반대 방향으로 이동하는 단어의 움직임을 관찰 할 수 있었다. 학습 기간과 테스트 기간에서 단어들의 동일한 방향으로 움직임이 많을수록 예측력이 높다 판단 하였고, 이는 연구의 방향이 하나의 방향으로 집중된다 예상하였다.

본 연구는 전문적인 정보를 담은 대량의 문서에서 주요 주제를 찾아내는 방법으로 기존 비지도 토픽 모델링 기법(LDA)에서 나아가 LDA 에 단어 임베딩 정보를 결합한 LDA2vec 을 활용하여 분석하였다. 학습데이터의 경우 바이오 관련 연구 15 만 건을 활용하여 임베딩 성능을 높이고자 하였으며, 본 프리트레인된 임베딩을 활용하여 준 지도 방식으로 LDA2vec 분석을 진행하였다. 본 연구에서 활용한 데이터 세트의 경우 딥러닝 방식으로 처음 분석이 진행되어, 지도 학습처럼

레이블링 데이터를 보유하고 있지 않았다. 이에 보건산업진흥원, 미래창조과학부, 사설 연구소에서 발간하는 신약 개발 동향 보고서로 성능을 판단할 수 밖에 없었고, 일부 전문적인 내용은 각 연구 분야 전문가만이 인식 할 수 있는 수준으로 분석의 어려움이 있었다. 하지만, 단어 임베딩과 토픽 모델링을 통해 상세 주제의 탐색 가능성을 제시하였다고 생각한다.

추후 연구의 진행 방향은 분석 이후의 수치화된 결괏값으로 그래픽 화를 통하여, 분석력과 전달력을 높이고자 한다. 또한 단어 임베딩을 넘어 최신 문장 임베딩 학습 방법인 ELMo(Embeddings from Language Models), BERT(Bidirectional Encoder Representations from Transformer)를 활용하여, 단어 임베딩에서 Transformer 방식으로 전환했을 때, 성능 향상과 분석 모델링에 대한 확인이 필요하다. 나아가 본 분석 방법은 학문의 연구 방향 설정뿐만 아니라, 신약 개발의 신물질 탐색, 국가 정책 분석 등 다양한 분야에서 활용의 가능성이 있다. 특히 기업체의 문자로 축척 된 VOC(Voice of Customer) 및 시장 동향 정보의 상세 분석에 이바지하여 효율적인 의사 결정을 할 수 있는 비즈니스 인텔리전스에 도움이 되고자 한다.

참 고 문 헌

1. Mostafa, J., W.J.I.P. Lam, and Management, *Automatic classification using supervised learning in a medical document filtering application.* 2000. **36**(3): p. 415-444.

2. Anto, S., *Supervised Machine Learning Approaches for Medical Data Set Classification-A Review 1.* 2011.

3. Forbes. *How Much Data Do We Creat Every Day?* 2018; Available from: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#23ac655760ba.

4. Intezari, A. and S.J.J.o.K.M. Gressel, *Information and reformation in KM systems: big data and strategic decision-making.* 2017. **21**(1): p. 71-91.

5. Koleck, T.A., et al., *Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review.* 2019. **26**(4): p. 364-379.

6. Hinton, G.E. and R.R.J.s. Salakhutdinov, *Reducing the dimensionality of data with neural networks.* 2006. **313**(5786): p. 504-507.

7. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* 2018.

8. Tang, D., et al. *User modeling with neural network for review rating prediction*. in *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.

9. Tshitoyan, V., et al., *Unsupervised word embeddings capture latent knowledge from materials science literature.* 2019. **571**(7763): p. 95.

10. 미래창조과학부, *생명공학백서*. 2015.

11. 보건산업진흥원, *주요국의 신약재창출(drug repositioning)동향과 전망*. 2014: 보건산업정보통계센터.

12. Hinton, G., J. McClelland, and D. Rumelhart, *Distributed representations. In the PDP Research Group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1. Foundations (pp. 77-109)*. 1986, London: MIT Press.

13. Rumelhart, D.J.L., *GE H inton, and RJ Williams.* 1986.

14. Bengio, Y., et al., *A neural probabilistic language model.* 2003. **3**(Feb): p. 1137-1155.

15. Mikolov, T., et al. *Recurrent neural network based language model*. in *Eleventh annual conference of the international speech communication association*. 2010.

16. Harris, Z., *Distributional structure. Word, 10 (2-3): 146-162. Reprinted in Fodor, J. A and Katz, JJ (eds.), Readings in the Philosophy of Language*. 1954, Englewood Cliffs, NJ: Prentice-Hall.

17. Pennington, J., R. Socher, and C. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

18. Mikolov, T., et al., *Efficient estimation of word representations in vector space.* 2013.

19. Levy, O., Y. Goldberg, and I.J.T.o.t.A.f.C.L. Dagan, *Improving distributional similarity with lessons learned from word embeddings.* 2015. **3**: p. 211-225.

20. Mikolov, T., et al. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.

21. Mikolov, T., W.-t. Yih, and G. Zweig. *Linguistic regularities in continuous space word representations*. in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013.

22. Bojanowski, P., et al., *Enriching word vectors with subword information.* 2017. **5**: p. 135-146.

23. 남춘호, *일기자료 연구에서 토픽모델링 기법의 활용가능성 검토.* 2016.

24. Salton, G. and M.J. McGill, *Introduction to modern information retrieval*. 1983: mcgraw-hill.

25. Deerwester, S., et al., *Indexing by latent semantic analysis.* 1990. **41**(6): p. 391-407.

26. Hofmann, T. *Probabilistic latent semantic analysis*. in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. 1999. Morgan Kaufmann Publishers Inc.

27. Blei, D.M., A.Y. Ng, and M.I.J.J.o.m.L.r. Jordan, *Latent dirichlet allocation.* 2003. **3**(Jan): p. 993-1022.

28. Moody, C.E.J.a.p.a., *Mixing dirichlet topic models and word embeddings to make lda2vec.* 2016.

29. Das, R., M. Zaheer, and C. Dyer. *Gaussian lda for topic models with word embeddings*. in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015.

30. Wang, Z., L. Ma, and Y. Zhang. *A hybrid document feature extraction method using latent Dirichlet allocation and word2vec*. in *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*. 2016. IEEE.

31. Bouma, G.J.P.o.G., *Normalized (pointwise) mutual information in collocation extraction.* 2009: p. 31-40.

32. Van de Cruys, T. *Two multivariate generalizations of pointwise mutual information*. in *Proceedings of the Workshop on Distributional Semantics and Compositionality*. 2011. Association for Computational Linguistics.

33. Newman, D., et al. *Automatic evaluation of topic coherence*. in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010. Association for Computational Linguistics.

34. Röder, M., A. Both, and A. Hinneburg. *Exploring the space of topic coherence measures*. in *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015. ACM.

35. Xu, X., et al., *Better conversations by modeling, filtering, and optimizing for coherence and diversity.* 2018.

36. R&D정보센터, *첨단 바이오의약품 산업전망과 신약개발 국내외 연구동향 분석*. 2018: 지식산업정보원.

Abstract

Unsupervised Word Embedding Based Topic Modeling Extracts Latent Biomedical Knowledge from Korean Gov. Research Proposals

Ahn, Byung Eun
Department of Engineering Practice
Graduate School of Engineering Practice
Seoul National University

Extracting latent knowledge from overwhelming text data has been a great challenge in the field of natural language processing. In recent years, with the significant improvement of deep learning, NLP is also experiencing a breakthrough and renewing the top score of various language tasks. However, to this date, understanding information, and creating new knowledge are human intelligence territory.

In this study, topic modeling algorithms were applied to extract latent information from massive documents (Korean government study proposals from 2006 to 2017, focusing on new drug development researches). To enhance domain-specific vocabulary detection during a preprocessing, NPMI(Normalized Pointwise Mutual Information) merged two nouns as a compound noun. For a topic modeling, unsupervised machine learning algorithm, LDA (Latent Dirichlet Allocation) was used to explore overall topic distributions. Furthermore, LDA2vec, which is a semi-supervised

deep learning model that training topic vectors along word embedding vectors in the same dimension, was applied to observe specific words correlation in a topic.

Without any labeling data and insertion of biochemical information, word embedding vectors that trained with topic vectors provide further interpretable information. Also, proposed three novel ways extract latent features from words and topics, and observe future tendency of research. These findings empathize the possibility of precise knowledge understanding. The expected field of applications are new drug investigation, business intelligence, and a variety of text data.

Keywords : Word Embedding, Semi Supervised Topic Modeling. LDA2vec, New Drug Investigation, NLP, Business Intelligence.

Student Number : 2018-29527

Appendix

기간별 LDA2vec 구현 결과

토픽	단어	단어	단어	단어	단어
1	KFDA,	Deacetylase,	예측시스템,	예비독성,	영상진단,
	병용처리,	LDL,	결핵백신,	소프트웨어,	apolipoprotein
2	NGS,	이질화,	형광단백질,	Epitope,	형질전환체,
	의료기,	생체모사,	transgenic,	Zebrafish,	dermal
3	<UNK>,	신경펩타이드,	알칼리,	염기서열,	노화방지,
	Fusion,	난치암,	Oncogene,	재활성화,	의료용
4	바이오산업,	올리고뉴클레오티	인공지능,	디지털,	항우울제,
	LPC,	피드백,	plaque,	대뇌피질,	폐렴구균
5	동물용,	청소년기,	repebo,	암세포주,	MCMT,
	Quality,	의료산업,	가톨릭대,	산학연,	연결고리
6	규격화,	paclitaxel,	출산율,	경조직,	의료복지,
	diabetes,	암유전자,	황색포도상구균,	데이터셋,	표면항원
7	세포신호전달체계	소재화,	XRD,	dynamics,	HCC,
	결정성,	약물수송,	melanogenesis,	기작연구,	CMO
8	영상진단,	JNK,	항당뇨,	ient,	sodium,
	chromatography	환경요인,	품질규격,	지재권,	Training
9	in-vitro,	Imatinib,	흡입기,	생체재료,	예후인자,
	sensitizer,	생쥐모델,	기대수명,	완제품,	산학연
10	약물전달체,	산학협력,	spike,	대한민국,	테라그노시스,
	multiple,	임상검증,	isomerase,	중소기업청,	스크린
11	대동물,	디프테리아,	예비독성,	conformation,	polyprotein,
	LPL,	미생물학,	의료복지,	repair,	transduction
12	뇌기능,	스쿠알렌,	항종양,	conformation,	유용물질,
	NSAID,	한약제제,	내분비,	소라페닙,	발암기전
13	Modeling,	생존인자,	생물자원,	암억제,	재활용,
	FXR,	Docking,	Prodrug,	NF-kB,	친환경
14	analyzer,	임상효과,	<UNK>,	식품의약,	tsutsu,
	단핵세포,	급성독성,	항진균,	면역독성,	상호결합
15	퇴행성질환,	CRISPR/Cas9,	ston,	발현양상,	사망자,
	적혈구,	drew,	homolog,	종양줄기세포,	product
16	초파리,	PGE2,	신경신호전달,	Surface,	plur,
	죽상경화증,	대량배양,	중합효소,	AAV,	나노기술
17	conjugated,	MMP,	management,	난치암,	Purity,
	프로모터,	뇌경색,	ozyme,	규격화,	element
18	stroke,	Collagen,	National,	수화물,	비타민D,
	<UNK>,	neuron,	가속기,	homolog,	CJD
19	막전압,	유도신경줄기세포	섬유소,	제조업체,	뇌신경질환,
	항진균,	OAB,	콘택트,	Microglia,	Ingredient
20	합토글,	림프관,	Artesu,	파트너,	여성암,
	병용처리,	세포주기,	백색지방,	분자역학,	후생유전체
21	경화제,	systemic,	보건학,	공중보건학,	exogenous,
	염증억제,	pocket,	인실리코,	MMR,	PLGA
22	독성작용,	표적발굴,	brogen,	한국형,	대동물,
	monoclo,	genomic,	줄기성,	IBD,	Flavonoid

LDA2vec 기간 4 결과

토픽	단어	단어	단어	단어	단어
1	리서치,	라이센스,	신경조직,	공통점,	고분해능,
	genome,	<UNK>,	chiral,	GABA,	value
2	인력양성,	chime,	ATG,	관상동맥질환,	친환경적,
	HPLC,	관절질환,	항염증제,	안토시아닌,	공정개선
3	lipid,	활성산소종,	인력양성,	2B4,	예방효과,
	세포손상,	화장품,	접근성,	HDL,	전자현미경
4	컴포넌트,	녹십자,	스케일업,	시냅스낭,	결정성,
	식품의약,	살모넬라균,	Smart,	HIV,	보툴리눔
5	transporter,	그리드,	컴퓨팅,	약물대사,	탄수화물,
	HBV,	동물시험,	프로젝트,	B형간염바이러스	화학유전학
6	genome,	폐암세포,	추출방법,	독성인자,	세포분열,
	공중합체,	연구윤리,	cytoplasm,	Guideline,	무분별
7	methylation,	영장류,	MAPK,	장기생존,	glycolysis,
	PAI-1,	약효평가,	흡수속도,	adipokine,	CIA
8	염기서열,	전자현미경,	신호체계,	단핵구,	표적유전자,
	homeostasis,	chromosome,	온도감응성,	Real-time,	methylation
9	선량분포,	피부재생,	요로감염,	발현량,	원인인자,
	유기물,	propo,	스크린,	스케일업,	화장품
10	genetic,	뇌신경질환,	생체조직,	estrogen,	driver,
	접근성,	phosphatid,	methylation,	syntheti,	extraction
11	포르말린,	췌장염,	세포반응,	혈중농도,	신장염,
	유방암세포,	CFA,	대국민,	표식인자,	계산과학
12	Selective,	소기관,	replication,	면역결핍,	사망자,
	분석시스템,	보건복지부,	백시니아,	제브라피쉬,	세포치료
13	genome,	oxidative,	efficient,	IRS-1,	유전적,
	세포학,	항산화제,	인력양성,	가동화,	solut
14	임피던스,	후유증,	Tumor,	lipid,	procedure,
	인터넷,	연구항,	PPAR-γ,	Survivin,	전기화학적
15	<UNK>,	당전이,	제조업,	digital,	인력양성,
	골격근,	컴퓨팅,	진통효과,	소아기,	중재연구
16	<UNK>,	뇌신경질환,	인력양성,	폐경기,	단백결합,
	Selective,	Smart,	liposome,	roughput,	항산화효소
17	Real-time,	형질전환마우스,	enhancer,	표식인자,	replication,
	간기능,	조직손상,	의과대학,	항원성,	식품의약
18	줄기세포치료,	치조골,	병리기전,	유전체연구,	치료기전,
	OLED,	환경요인,	민간요법,	glycolysis,	인력양성
19	치료기전,	인력양성,	전자현미경,	시판허가,	이온통로,
	분자량,	신경퇴행성,	Medical,	선도화합물,	항균효과

LDA2vec 기간 3 결과

토픽	단어	단어	단어	단어	단어
1	Biological,	gold,	장기이식,	aldosteron,	fluorouracil,
	성인병,	면역력,	전립선,	tubulin,	류머티즘
2	HCS,	유기화합물,	중소기업,	질병진단,	생명과학,
	친환경적,	기전규명,	진통효과,	meter,	파이프라인
3	Bioreactor,	검출기술,	pool,	Kinase,	치료약제,
	인간배아줄기세포	전자현미경,	트립토판,	혈청학,	QPCR
4	세포기능,	이온채널,	면역조절,	면역조절기능,	spinal,
	NF-κB,	product,	뇌기능,	COX-2,	검출기술
5	배당체,	DMA,	polar,	항암작용,	MTT,
	플라보노이드,	약용식물,	수산기,	dynamics,	유방암
6	추출공정,	폐암세포,	product,	유기화합물,	Global,
	Scale-up,	코호트,	신경퇴행성,	photo,	전립선
7	예방효과,	VEGF-A,	거버넌스,	HPLC,	배큘로바이러스,
	cathepsin,	생체시료분석,	toxicity,	detector,	angiogenesis
8	x-ray,	seque,	multi,	product,	inhibitors,
	cytokine,	valent,	결합부위,	hydrogenase,	strip
9	리소그래피,	erefore,	improve,	plaque,	alcohol,
	chromatin,	chaperone,	needed,	대사관,	치료약제
10	inhibitors,	항균성,	MRSA,	항생물질,	VRE,
	곰팡이,	유전학,	생식독성,	비만치료,	질병발생
11	연구동향,	의과대학,	일반인,	분류학,	HepG2,
	대사효소,	organic,	polar,	극소화,	류머티즘
12	결합단백질,	유기화합물,	원천기술개발,	mammal,	QSAR,
	zinc,	질환치료제,	PKC,	injection,	치료표적
13	발생학,	흡광도,	photo,	deacetylase,	Histone,
	internalization,	원심분리,	migration,	생명공학기술,	embryo
14	다중표적,	opioid,	multi-target,	경피흡수,	재형성,
	동맥경화반,	해상도,	동위원소,	클러스터,	핵의학
15	일반인,	측정기술,	섬유모세포,	CD4,	표식자,
	치료제개발,	paclitaxel,	VSIG4,	친환경적,	toxicity
16	치료표적,	암치료법,	미개척,	알고리즘,	치료기술개발,
	해상도,	metalloprotein,	의료기기,	plastic,	영상유도
17	product,	심포지엄,	연구동향,	toxicity,	LC-MS,
	alcohol,	유전공학,	Scale-up,	인체세포,	detector
18	핵의학,	영양소,	난치성질환,	purification,	치료약제,
	보조제,	기전규명,	product,	통증완화,	erie
19	세포기능,	검출기술,	mediator,	치료약제,	Double,
	multi,	합성물질,	seque,	미세유체,	영양소

LDA2vec 기간 2 결과

토픽	단어	단어	단어	단어	단어
1	<UNK>,	광증감제,	neuron,	백내장,	히알루론산,
	Phage,	epithelium,	solut,	drugs,	organic
2	골격계,	Oral,	신호조절,	국제특허,	혈관세포,
	GPCRs,	약물대사,	실험법,	overy,	응용기술
3	대기업,	lori,	오징어,	공정개발,	항노화,
	골관절염,	항산화물질,	Helicobacter,	bioreactor,	활성산소종
4	생명체,	항섬유화,	heterogeneity,	ectro,	MAP,
	연구윤리,	pept,	고혈당,	PSMA,	생체분자
5	식품의약품안전청	파이프라인,	roxy,	KFDA,	국소부위,
	보습력,	stimulator,	종양성장,	무혈청,	정보제공
6	골격계,	CMC,	유전자발현,	신호체계,	autophagy,
	혈관생성,	재형성,	종양성장,	유기금속,	상호관련성
7	업그레이드,	histology,	Inhibitor,	면역질환,	항히스타민제,
	liposome,	골격계,	피부자극,	보편화,	interface
8	약물후보물질,	신호전달기전,	골격계,	미세구조,	AMPA,
	뇌허혈,	CCR,	신호물질,	depression,	artery
9	어린이,	노약자,	면역성,	탐식세포,	측정기술,
	항진균제,	uct,	미생물,	rally,	Mock
10	typing,	exam,	상호관련성,	유비쿼터스,	표면단백질,
	교육자,	romo,	Biomarker,	식습관,	CAD
11	플랫폼,	의료기기,	연구윤리,	콘텐츠,	일반인,
	해외시장,	PARP-1,	항노화,	세포분열,	무독성
12	TIMP,	neuron,	면역질환,	나노기술,	국산화,
	ckgrou,	제대혈,	중성지방,	이상지질혈증,	간질환
13	활성인자,	E-cadherin,	생명체,	암발생,	export,
	TGF-β,	비용절감,	두경부,	간질환,	골격근
14	업그레이드,	저해활성,	박테리아,	anti-inflammato	단백질발현,
	배양체,	구강암,	후반기,	spacer,	이차대사산물
15	세포분화,	생명체,	세포분열,	neuron,	종양인자,
	p53,	전이금속,	parent,	칼슘채널,	복합물
16	상호관련성,	골격계,	카나비노이드,	cisplatin,	비교유전체,
	발현양상,	광증감제,	산화환원효소,	변이유전자,	노약자
17	EMT,	TGF-β,	IL-6,	genomic,	장기이식,
	event,	기능회복,	homolog,	섬유모세포,	network

LDA2vec 기간 1 결과